

SNIFE for Memory-Limited PCA From Incomplete Data

Armin Eftekhari, Laura Balzano, Dehui Yang, Michael B. Wakin*

December 6, 2016

Abstract

The linear subspace model is pervasive in science and engineering and particularly in large datasets which are often incomplete due to missing measurements and privacy issues. Therefore, a critical problem in modeling is to develop algorithms for estimating a low-dimensional subspace model from incomplete data efficiently in terms of both computational complexity and memory storage. In this paper we study an algorithm that processes blocks of incomplete data to estimate the underlying subspace model. Our algorithm has a simple interpretation as optimizing the subspace to fit the observed data block but remain close to the previous estimate. We prove a linear rate of convergence for the algorithm and our rate holds with high probability.

1 Introduction

Linear models are the backbone of modern sciences and *principal component analysis* (PCA) has traditionally been an indispensable tool in studying collected data [1, 2, 3, 4]. In dimensionality reduction, for example, PCA searches for the linear model that best describes the data [5]. In this work we are particularly interested in applying PCA to data that suffers from erasures, while limited storage is available. In recommender systems, for instance, data is highly incomplete and yet so massive that can only be processed in small chunks [6].

To focus our efforts and as our guiding example throughout, consider incomplete data from an unknown subspace, presented sequentially to the user who, due to hardware limitations, can only store small amounts of data. We are here interested in developing a *streaming* algorithm for PCA from incomplete measurements.

More concretely, consider an r -dimensional subspace \mathcal{S} with orthonormal basis $S \in \mathbb{R}^{n \times r}$. For an integer T , let the coefficient vectors $\{q_t\}_{t=1}^T \subset \mathbb{R}^r$ be independent copies of a random vector $q \in \mathbb{R}^r$. At time $t \in [1 : T] := \{1, 2, \dots, T\}$, we observe each entry of $s_t := S \cdot q_t \in \mathcal{S}$ independently with a probability of p , and we collect the measurements in $y_t \in \mathbb{R}^n$, supported on a random index set $\omega_t \subseteq [1 : n]$. Formally, we will write this measurement process as $y_t = P_{\omega_t}(s_t) = P_{\omega_t} \cdot s_t$, where $P_{\omega_t} \in \mathbb{R}^{n \times n}$ is the projection onto the coordinate set ω_t , i.e. it equals one on its diagonal entries corresponding to the index set ω_t , and is zero elsewhere.

Our objective is to identify the subspace \mathcal{S} from the measurements $\{y_t\}_{t=1}^T$ supported on the index sets $\{\omega_t\}_{t=1}^T$. Assuming that $r = \dim(\mathcal{S})$ is known *a priori* (or estimated from data by other means), we present the SNIFE algorithm for this task in Section 2 and, in particular, provide helpful insight about the algorithm in Section 2.1. As summarized in Section 3, SNIFE converges, globally and linearly, to the true subspace under reasonable requirements. A detailed review of prior art is deferred to Section 4. The performance of SNIFE and rival algorithms are examined numerically in Section 5, where we find that SNIFE improves upon the state of the art. Technical proofs appear in Section 6 and in the appendices, with Appendix A (Toolbox) collecting some of the frequently used mathematical tools.

Before delving into the details, let us conclude this section with an example. Suppose that $\mathcal{S} \subset \mathbb{R}^{1000}$ is a generic subspace of dimension $r = 3$ and take $p = 0.1$. Then, SNIFE produces a sequence of estimates of \mathcal{S} ; the estimation error (with a metric to be specified shortly) versus t is plotted in Figure 1.

*AE is with the Alan Turing Institute in London. LB is with the Department of Electrical Engineering and Computer Science at the University of Michigan, Ann Arbor. DY and MBW are with the Department of Electrical Engineering and Computer Science at Colorado School of Mines. (E-mails: aeftekhari@turing.ac.uk; girasole@umich.edu; dyang@mines.edu; mwakin@mines.edu)

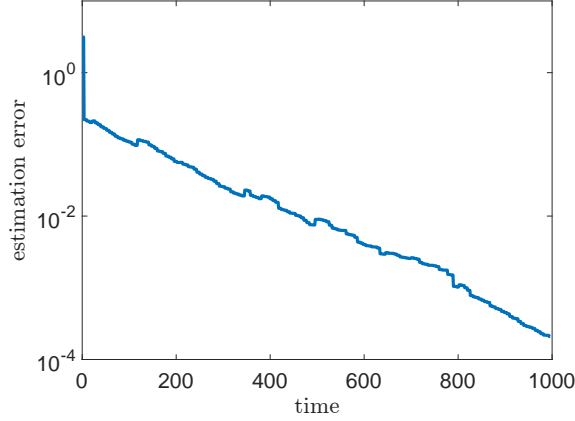


Figure 1: This paper introduces SNIPE for memory-limited PCA from incomplete data. This figure shows the estimation error of SNIPE versus time in recovering a generic 3-dimensional subspace from data subsampled by a factor of 10, received sequentially. The algorithm and the error metric are described below.

2 SNIPE

In this section, we propose Subspace Navigation via Interpolation from Partial Entries (SNIPE), an algorithm for subspace identification from incomplete data, received sequentially.

To formally present SNIPE, let us introduce some additional notation. Recall the incoming sequence of measurement vectors $\{y_t\}_{t=1}^T \subset \mathbb{R}^n$, supported on the index sets $\{\omega_t\}_{t=1}^T \subseteq [1 : n]$. For an integer K , fix block sizes $\{b_k\}_{k=1}^K$ such that $\sum_{k=1}^K b_k = T$ and $b_k \geq r$ for all k . We partition the data into K (non-overlapping) blocks $\{Y_k\}_{k=1}^K$, where $Y_k \in \mathbb{R}^{n \times b_k}$ for each k .

At a high level, SNIPE processes the first measurement block Y_1 to produce an estimate $\hat{\mathcal{S}}_1$ of the true subspace \mathcal{S} . This estimate is then iteratively updated after receiving each of the new blocks $\{Y_k\}_{k=2}^K$, thereby producing a sequence of estimates $\{\hat{\mathcal{S}}_k\}_{k=2}^K$. Every $\hat{\mathcal{S}}_k$ is an r -dimensional subspace of \mathbb{R}^n with orthonormal basis $\hat{S}_k \in \mathbb{R}^{n \times r}$; the particular choice of orthonormal basis is inconsequential throughout the paper.

More concretely, SNIPE sets $\hat{\mathcal{S}}_1$ to be the span of the top r left singular vectors of Y_1 , namely the left singular vectors corresponding to the largest r singular values of Y_1 , with ties broken arbitrarily. Then, at iteration $k \in [2 : K]$ and given the previous estimate $\hat{\mathcal{S}}_{k-1} = \text{span}(\hat{S}_{k-1})$, SNIPE processes the columns of the k th measurement block Y_k by forming the matrix

$$R_k = \begin{bmatrix} \cdots & y_t + P_{\omega_t^c} \hat{S}_{k-1} (P_{\omega_t} \hat{S}_{k-1})^\dagger y_t & \cdots \end{bmatrix} \in \mathbb{R}^{n \times b_k}, \quad \sum_{k'=1}^{k-1} b_{k'} + 1 \leq t \leq \sum_{k'=1}^k b_{k'}, \quad (1)$$

where $P_{\omega_t^c} = I_n - P_{\omega_t} \in \mathbb{R}^{n \times n}$ projects a vector onto the complement of the index set ω_t . SNIPE then updates its estimate by setting $\hat{\mathcal{S}}_k$ to be the span of the top r left singular vectors of R_k . Figure 2 summarizes these steps.

2.1 Interpretation of SNIPE

SNIPE has a natural interpretation as a solver for a non-convex optimization program, which we next discuss. First, however, let us enrich our notation: Recall the measurement blocks $\{Y_k\}_{k=1}^K$ and let the (random) index set $\Omega_k \subseteq [1 : n] \times [1 : b_k]$ be the support of Y_k , for every k . We write that $Y_k = P_{\Omega_k}(S_k)$, where the columns of $S_k \in \mathbb{R}^{n \times b_k}$ belong to the true subspace \mathcal{S} , namely $S_k = S \cdot Q_k^*$ for some coefficient matrix $Q_k \in \mathbb{R}^{b_k \times r}$. Here, $P_{\Omega_k}(S_k)$ retains only the entries of S_k on the index set Ω_k , setting the rest to zero. Note that $\{S_k\}_{k=1}^K$ and $\{Q_k\}_{k=1}^K$ are formed by partitioning $\{s_t\}_{t=1}^T$ and $\{q_t\}_{t=1}^T$ into K blocks, respectively. Likewise, $\{\Omega_k\}$ are formed from $\{\omega_t\}$.

Input:

- Dimension r ,
- Received data $\{y_t\}_{t=1}^T \subset \mathbb{R}^n$ supported on index sets $\{\omega_t\}_{t=1}^T \subseteq [1 : n]$,
- Number of blocks K and block sizes $\{b_k\}_{k=1}^K$, where $\sum_{k=1}^K b_k = T$ and $b_k \geq r$ for all k .

Output: r -dimensional subspace \hat{S}_K .

Body:

- Form $Y_1 \in \mathbb{R}^{n \times b_1}$ from the first b_1 measurement vectors $\{y_t\}_{t=1}^{b_1}$. Let \hat{S}_1 , with orthonormal basis $\hat{S}_1 \in \mathbb{R}^{n \times r}$, be the span of the top r left singular vectors of Y_1 , namely those corresponding to the largest singular values. Ties are broken arbitrarily.
- For $k \in [2 : K]$, repeat:
 - Set $R_k \leftarrow \{\}$.
 - For $t = \sum_{k'=1}^{k-1} b_{k'} + 1 : \sum_{k'=1}^k b_{k'}$, repeat
 - * Set

$$R_k \leftarrow \begin{bmatrix} R_k & y_t + P_{\omega_t^c} \hat{S}_{k-1} (P_{\omega_t} \hat{S}_{k-1})^\dagger y_t \end{bmatrix},$$
 where $P_{\omega_t} \in \mathbb{R}^{n \times n}$ equals one on its diagonal entries corresponding to the index set ω_t , and is zero elsewhere. Likewise, $P_{\omega_t^c}$ projects a vector onto the complement of the index set ω_t .
 - Let \hat{S}_k , with orthonormal basis $\hat{S}_k \in \mathbb{R}^{n \times r}$, be the span of the top r left singular vectors of R_k . Ties are broken arbitrarily.
- Return \hat{S}_K .

Figure 2: SNIPE for memory-limited PCA from incomplete data.

With this introduction, let us form

$$Y := \begin{bmatrix} Y_1 & Y_2 & \cdots & Y_K \end{bmatrix} \in \mathbb{R}^{n \times T}, \quad (2)$$

by concatenating all measurement blocks, supported on $\Omega \subseteq [1 : n] \times [1 : T]$. To find the true subspace \mathcal{S} , one might then solve

$$\begin{cases} \min_{\mathcal{U}, X} \|P_{\mathcal{U}^\perp} X\|_F^2, \\ P_\Omega(X) = Y, \end{cases} \quad (3)$$

where the minimization is over all r -dimensional subspaces $\mathcal{U} \subset \mathbb{R}^n$ and matrices $X \in \mathbb{R}^{n \times T}$. Also, $P_{\mathcal{U}^\perp} \in \mathbb{R}^{n \times n}$ is the orthogonal projection onto the orthogonal complement of subspace \mathcal{U} , and $P_\Omega(X)$ retains only the entries of X on the index set Ω , setting the rest to zero. In particular, with full measurements ($\Omega = [1 : n] \times [1 : T]$), Program (3) reduces to PCA, as it returns $X = Y$ and sets \mathcal{U} equal the span of the top r left singular vectors of Y . Note also that Program (3) is a non-convex problem because the Grassmannian $\mathbb{G}(n, r)$, the set of all r -dimensional subspaces in \mathbb{R}^n , is a non-convex set.¹

Instead of Program (3), consider the equivalent program

$$\begin{cases} \min \sum_{k=1}^K \|P_{\mathcal{U}_k}^\perp X_k\|_F^2, \\ P_{\Omega_k}(X_k) = Y_k, \quad k \in [1 : K], \\ \mathcal{U}_1 = \mathcal{U}_2 = \cdots = \mathcal{U}_K, \end{cases} \quad (4)$$

where the minimization is over all subspaces $\{\mathcal{U}_k\} \subset \mathbb{G}(n, r)$ and matrices $\{X_k\}$, with $X_k \in \mathbb{R}^{n \times b_k}$ for every k .

Consider the following approximate solver for Program (4):

- Setting $X_1 = Y_1$ in Program (4), we minimize $\min \|P_{\mathcal{U}_1}^\perp Y_1\|_F^2$ over $\mathcal{U}_1 \in \mathbb{G}(n, r)$ and find a minimizer to be the span of top r left singular vectors of Y_1 , namely $\widehat{\mathcal{S}}_1$ in SNIPE.
- For $k \in [2 : K]$, repeat the following:
 - Setting $\mathcal{U}_k = \widehat{\mathcal{S}}_{k-1}$ in Program (4), we solve

$$\begin{cases} \min \|P_{\widehat{\mathcal{S}}_{k-1}^\perp} X_k\|_F^2, \\ P_{\Omega_k}(X_k) = Y_k, \end{cases} \quad (5)$$

over all matrices $X_k \in \mathbb{R}^{n \times b_k}$. In Appendix B, we verify that the minimizer of Program (5) is R_k in SNIPE.

- Setting $X_k = R_k$ in Program (4), we minimize $\min \|P_{\mathcal{U}_k}^\perp R_k\|_F^2$ over $\mathcal{U}_k \in \mathbb{G}(n, r)$ to find $\widehat{\mathcal{S}}_k$ in SNIPE.

Note that following this procedure precisely produces $\{R_k\}$ and $\{\widehat{\mathcal{S}}_k\}$ in SNIPE.

Another insight about the choice of R_k in (1) is as follows. At the beginning of the k th iteration of SNIPE with $k \geq 2$, the available estimate of the true subspace is $\widehat{\mathcal{S}}_{k-1}$ with orthonormal basis $\widehat{\mathcal{S}}_{k-1}$. Given a new measurement vector $y \in \mathbb{R}^n$, supported on the index set $\omega \subseteq [1 : n]$, $z = \widehat{\mathcal{S}}_{k-1}(P_\omega \widehat{\mathcal{S}}_{k-1})^\dagger y$ best approximates y in $\widehat{\mathcal{S}}_{k-1}$ (in the ℓ_2 sense). In order to agree with the measurements, we minimally adjust this to $y + P_{\omega^c} z$, where P_{ω^c} projects onto the complement of ω . This indeed matches the expression for the columns of R_k in SNIPE. We note that this type of “least-change” strategy has been successfully used in the development of quasi-Newton methods for optimization [7, Chapter 6].

Lastly, it would be interesting to find a statistical interpretation for (1), perhaps in the form of a conditional expectation on the new measurement block. This might be a nontrivial task, however, as it seems to require forming a prior distribution on the Grassmannian.

¹The Grassmannian can be embedded in $\mathbb{R}^{n \times n}$ via the map that takes $\mathcal{U} \in \mathbb{G}(n, r)$ to the orthogonal projection $P_{\mathcal{U}}$. The resulting submanifold of $\mathbb{R}^{n \times n}$ is a non-convex set.

3 Performance of SNIPE

This section summarizes the theoretical guarantees for SNIPE, with the derivations deferred to Section 6 and the appendices. To measure the performance of SNIPE, we naturally use principal angles as our error metric. More specifically, recall that \mathcal{S} and $\hat{\mathcal{S}}_K$ denote the true subspace and the output of SNIPE, respectively. Then, the i th largest singular value of $P_{\mathcal{S}^\perp} P_{\hat{\mathcal{S}}_K}$ equals $\sin(\theta_i(\mathcal{S}, \hat{\mathcal{S}}_K))$, where

$$\theta_1(\mathcal{S}, \hat{\mathcal{S}}_K) \geq \theta_2(\mathcal{S}, \hat{\mathcal{S}}_K) \geq \cdots \geq \theta_r(\mathcal{S}, \hat{\mathcal{S}}_K)$$

denote the principal angles between \mathcal{S} and $\hat{\mathcal{S}}_K$ [8]. Our error metric is then

$$d_{\mathbb{G}}(\mathcal{S}, \hat{\mathcal{S}}_K) := \sqrt{\frac{1}{r} \sum_{i=1}^r \sin^2(\theta_i(\mathcal{S}, \hat{\mathcal{S}}_K))} = \frac{\|P_{\mathcal{S}^\perp} P_{\hat{\mathcal{S}}_K}\|_F}{\sqrt{r}}. \quad (6)$$

The concept of *coherence* is also critical in parsing our results, since we consider entrywise subsampling. The coherence of an r -dimensional subspace \mathcal{S} with orthonormal basis $S \in \mathbb{R}^{n \times r}$ is defined as

$$\eta(\mathcal{S}) := \frac{n}{r} \max_i \|S[i, :]\|_2^2, \quad (7)$$

where $S[i, :]$ is the i th row of S . It is easy to verify that $\eta(\mathcal{S})$ is independent of the choice of orthonormal basis S , and that $1 \leq \eta(\mathcal{S}) \leq \frac{n}{r}$. It is also common to say that \mathcal{S} is *coherent* (*incoherent*) when $\eta(\mathcal{S})$ is large (small). Loosely speaking, when \mathcal{S} is coherent, its orthonormal basis S is “spiky.” An example is when \mathcal{S} is the span of a column-subset of the identity matrix. In contrast, when \mathcal{S} is incoherent, entries of S tend to be “diffuse.” Not surprisingly, identifying a coherent subspace from subsampled data may require many more measurements [9, 10, 11].

Lastly, we will often use \lesssim and \gtrsim to suppress universal constants, as well as the standard big-O and big-Omega notations, $O(\cdot)$ and $\Omega(\cdot)$, respectively. With this setup, our main result is as follows.

Theorem 1. [Performance of SNIPE] *Consider an r -dimensional subspace \mathcal{S} with orthonormal basis $S \in \mathbb{R}^{n \times r}$. For an integer T , let the coefficient vectors $\{q_t\}_{t=1}^T \subset \mathbb{R}^r$ be independent copies of a random vector $q \in \mathbb{R}^r$. For every $t \in [1 : T]$, we observe each coordinate of $s_t = Sq_t \in \mathcal{S}$ independently and with a probability of p , and we collect the measurements in $y_t \in \mathbb{R}^n$, setting unobserved entries to zero.*

For an integer K , fix block sizes $\{b_k\}_{k=1}^K$ such that $\sum_{k=1}^K b_k = T$ and $b_k \geq r$ for all k . Fix also $\alpha \gtrsim 1, \nu \geq 1$ and $\{\eta_k\}_{k=1}^K$ such that $1 \leq \eta_k \leq b_k/(\alpha^2 \nu^2 r)$ for all k . By partitioning $\{q_t\}_t$, form $\{Q_k\}_{k=1}^K$ with $Q_k \in \mathbb{R}^{b_k \times r}$ for each k . Let $\nu(Q_k)$ be the condition number of Q_k , namely the ratio of its largest and smallest singular values. Set also $\mathcal{Q}_k = \text{span}(Q_k)$ with coherence $\eta(\mathcal{Q}_k)$ (see (7)).

Then, with $\hat{\mathcal{S}}_K$ being the output of SNIPE, it holds that

$$d_{\mathbb{G}}(\mathcal{S}, \hat{\mathcal{S}}_K) \lesssim 2^{-K} \cdot \alpha \cdot \nu \cdot \sqrt{\left(1 \vee \frac{n}{b_1}\right) \frac{(\eta_1 \vee \eta(\mathcal{S})) r \log(n \vee \max_k b_k)}{pn}}, \quad (8)$$

except with a probability of

$$K \cdot O\left(e^{-\alpha} + (\min_k b_k)^{-\alpha}\right) + \sum_{k=1}^K \Pr[\nu(Q_k) > \nu] + \Pr[\eta(\mathcal{Q}_k) > \eta_k], \quad (9)$$

and provided that

$$p \gtrsim \frac{\alpha^6 \nu^6}{\alpha^2 \nu^2 - C_1} \left(1 \vee \frac{n}{b_1}\right) \frac{(\eta_1 \vee \eta(\mathcal{S})) r^2 \log^4(n \vee \max_k b_k)}{n}, \quad p \geq 1 - \frac{C_1}{\alpha^2 \nu^2}, \quad (10)$$

for an absolute constant $C_1 > 0$. Above, $a \vee b = \max(a, b)$ for $a, b \in \mathbb{R}$.

A few remarks are in order, starting with a simplified statement of Theorem 1.

Remark 1. [Discussion] Let us first discuss the essence of Theorem 1, before making things more concrete in the next remarks. Suppose that $\eta(\mathcal{S}) = 1$, namely the true subspace is incoherent. Suppose also that we partially measure “generic” vectors drawn from \mathcal{S} and choose the following for the parameters in Theorem 1: $\alpha = O(1)$, $\nu = O(1)$, $b_1 = \Omega(n)$, $b_2 = b_3 = \dots = b_K = \Omega(r \log r)$, and $\eta_k = O(\log b_k)$ for all k .² In this case, Theorem 1 roughly states that the estimation error of SNIPE obeys

$$d_{\mathbb{G}}(\mathcal{S}, \hat{\mathcal{S}}_K) \lesssim 2^{-K} \log n \sqrt{\frac{r}{pn}},$$

with high probability, and provided that

$$p \gtrsim \frac{r^2 \log^5 n}{n}, \quad p \geq C_2,$$

for some constant $C_2 \in (0, 1)$. In other words, the estimation error of SNIPE decays exponentially fast provided we observe $\max(r^2 \log^5 n, C_2 n)$ entries of every incoming vector. In the following remarks, we will discuss different aspects of this result in more details.

Remark 2. [Global linear convergence] According to Theorem 1, SNIPE converges globally and linearly to the true subspace, with high probability and if the sampling probability p is sufficiently large. In this sense, Theorem 1 provides the first comprehensive analysis of a class of “least-change” algorithms that includes SNIPE and GROUSE [13], as discussed in more detail in Section 4.

Remark 3. [Sampling probability] A notable shortcoming of Theorem 1 is the requirement in (10) that $p \geq 1 - C_1/\alpha^2 \nu^2$. A perhaps less pressing issue is the quadratic (rather than linear) relation in (10) between the ambient dimension n and the typically small rank r .

As the strong numerical performance of SNIPE in Section 5 suggests, these drawbacks of Theorem 1 appear to be artifacts of our proof technique that have resisted all efforts to remove them. See Remark 8 for the technical roots of these difficulties.

Remark 4. [Coefficients] Influence of the coefficient vectors $\{q_t\}_{t=1}^T$ in Theorem 1, as captured by the coherence factors $\{\eta(Q_k)\}_{k=1}^K$, is often mild. For example, suppose that $\{q_t\}_{t=1}^T$ are independent copies of a standard Gaussian vector, in which case $\{Q_k\}_{k=1}^K$ would become independent standard random Gaussian matrices. Then, with fixed $\alpha \geq 1$ as in Theorem 1, familiar arguments in random matrix theory yield that

$$\nu(Q_k) = O(1), \quad \eta(Q_k) \lesssim \alpha \log b_k, \quad \forall k \in [1 : K], \quad (11)$$

when $\min_k b_k \gtrsim \alpha^2 r$ and except with a probability of at most $K(e^{-\alpha r} + (\min_k b_k)^{-\alpha})$. For the sake of completeness, (11) is proved in Appendix J. Therefore, with the choice of

$$\nu = O(1), \quad \eta_k = O(\alpha \log b_k), \quad \forall k \in [1 : K], \quad (12)$$

the failure probability in (9) reads

$$K \cdot O\left(e^{-\alpha} + (\min_k b_k)^{-\alpha}\right) + K \cdot \left(e^{-\alpha r} + (\min_k b_k)^{-\alpha}\right) = K \cdot O\left(e^{-\alpha} + (\min_k b_k)^{-\alpha}\right). \quad (13)$$

We close this remark by pointing out that the dependence on $\{\eta(Q_k)\}_k$ in Theorem 1 is not an artifact of the analysis. For example, suppose that each column of every Q_k contains only two nonzero entries, both equal to one, located at indexes $2i$ and $2i + 1$, with integer i selected uniformly at random. Then, for all k , $\eta(Q_k) = \frac{n}{2r}$ is large and, even when $p = 1$, it is impossible to correctly recover the true subspace \mathcal{S} .

Remark 5. [Block sizes] For the sake of discussion, suppose that the coefficient vectors are independent standard Gaussian vectors and take ν and $\{\eta_k\}_k$ as in Remark 4. Both (8) and (10) strongly suggest setting the first block size in SNIPE as $b_1 = \Omega(n)$. Remark 4 suggests setting the rest of block sizes as $b_k = \Omega(r)$, $k \geq 2$. The requirement on b_1 is likely *not* an artifact, as discussed in Remark 8; there appear to be pathological scenarios in which $b_1 = \Omega(n)$ is necessary for global convergence of SNIPE. However, in all simulations in Section 5, we set $b_1 = b_2 = \dots = b_K = \Omega(r)$ without facing any convergence problems.

²Strictly speaking, b_k must satisfy $r \lesssim b_k / \log b_k$ for $k \geq 2$, in light of Theorem 1 and Remark 4, specifically (12). This requirement can also be expressed by means of the *Lambert W-function* [12, §4.13]. However, $\log \log r$ is often a small constant, which justifies the simpler statement $b_k = \Omega(r \log r)$ in Remark 1.

Remark 6. [Computational complexity] We measure the algorithmic complexity of SNIPE by calculating the average number of floating-point operations performed on an incoming vector. Again, consider random Gaussian coefficients and the choice of ν and $\{\eta_k\}$ in Remark 4. Then, in light of Remark 5, let us take $b_1 = \Omega(n)$ and $b_k = \Omega(r)$ for $k \geq 2$.

Iteration $k = 1$ of SNIPE involves (approximately) finding the top r left singular vectors of $Y_1 \in \mathbb{R}^{n \times b_1}$ which, using a randomized SVD, could be done after $O(rn^2)$ operations or equivalently $O(rn)$ operations per vector [14, Section 1.6]. At the k th iteration with $k \geq 2$, SNIPE requires finding the pseudo-inverse of $P_{\omega_j} \hat{S}_{k-1} \in \mathbb{R}^{n \times r}$ for each incoming vector which costs $O(r^2n)$ operations. Overall, the computational complexity of SNIPE in iteration $k \geq 2$ is $O(r^2n)$ operations per vector. As further discussed in Section 4, this matches the complexity of other algorithms for memory-limited PCA from limited measurements.

Remark 7. [Storage] We measure the storage required by SNIPE by calculating the number of memory elements stored by SNIPE at any given instant. Again, consider random Gaussian coefficients and choice of ν and $\{\eta_k\}$ in Remark 4. Then, in light of Remark 5, let us take $b_1 = \Omega(n)$ and $b_k = \Omega(r)$ for $k \geq 2$.

The first iteration ($k = 1$) of SNIPE requires keeping $Y_1 \in \mathbb{R}^{n \times b_1}$ in the memory and, therefore, has a storage requirement of $O(pn^2)$ or simply $O(n^2)$ memory elements. However, the rest of the iterations have a far lower storage requirement of $O(rn)$ memory elements. Indeed, at the k th iteration for $k \geq 2$, SNIPE must store the current estimate $\hat{S}_{k-1} \in \mathbb{R}^{n \times r}$ and the new measurement block $Y_k \in \mathbb{R}^{n \times r}$. This translates into $O(rn) + O(prn)$ or simply $O(rn)$ memory elements. See Section 4 for comparison with other algorithms.

Remark 8. [Proof strategy and room for improvement] To prove Theorem 1, we first establish local linear convergence. That is, we establish in Lemma 3 that SNIPE converges to the true subspace \mathcal{S} (with high probability and when p is large enough) if the algorithm is initialized sufficiently close to \mathcal{S} , namely when \hat{S}_1 is close enough to \mathcal{S} . The proof of this claim crucially involves a tight perturbation bound (Lemma 6), as well as controlling the spectral norms of $P_{\mathcal{S}^\perp} R_k$ and $R_k - SQ_k^*$ (see (1)), which is difficult due to the nature of R_k . As a proxy for the spectral norm, we bound the Frobenius norm of these quantities in Appendix D and five ensuing appendices. This compromise, however, leads to the quadratic (rather than linear) dependence of p on r in (10), which is likely an artifact of the proof technique. Other key technical steps to obtain Lemma 3 include avoiding statistical dependence in Section D.2, as well as finding a tight bound on a key random variable in Lemma 12 by directly analyzing the associated “normal equation.” Another artifact of this argument is the constant lower bound on p , which appears difficult to remove.

The next step towards the proof of Theorem 1 is establishing in Lemma 2 that \hat{S}_1 , produced by SNIPE, is sufficiently close to \mathcal{S} . This lemma allows us to guarantee that \hat{S}_1 is a good initialization that would “activate” the local linear convergence in the rest of SNIPE iterations ($k \geq 2$) promised by Lemma 3. The proof of Lemma 2 involves a standard application of the matrix Bernstein inequality in Appendix C.

Lastly, we combine Lemmas 2 and 3 together to complete the proof. In particular, we find that, for \hat{S}_1 to be a good initialization, taking $b_1 = \Omega(n)$ is necessary in the proof. Whether this requirement is an artifact of the proof is not clear to the authors. In particular, it has been recently established that any local minimum of the objective function in (3) (after small changes) is also a global minimum [15]. However, it is not clear what the implication of this result is here since SNIPE does not belong to the class of solvers in Theorem 2.3 in [15]. Further investigating the requirement $b_1 = \Omega(n)$ is needed.

4 Related Work

Among several algorithms that have been proposed for tracking low-dimensional structure in a data set from partially observed streaming measurements [10, 16, 17, 18], SNIPE might be most closely related to GROUSE [13, 19]. GROUSE performs memory-limited PCA from incomplete data using stochastic gradient projection on the Grassmannian, updating its estimate of the true subspace with each new measurement vector.

Both GROUSE and SNIPE were designed based on the principal of least change, discussed in Section 2.1. In fact, when GROUSE is sufficiently close to the true subspace and with a specific choice of its step length, both algorithms have nearly identical updates (see [9, Equation 1.9]).

Local convergence of GROUSE, in expectation, was recently established in [9]. More specifically, [9] stipulates that, if the current estimate \hat{S}_k is sufficiently close to the true subspace \mathcal{S} , then \hat{S}_{k+1} will be even

closer to \mathcal{S} and this leads to a linear local convergence in expectation.

Our own update was inspired by that of GROUSE when we found zero-filled updates were unreliable. However, GROUSE was derived as a purely streaming algorithm, and it therefore is not designed to leverage common low-rank structure that may be revealed when a block of vectors is processed at once. Therefore for each block, SNIPE achieves a more significant reduction in error than is possible with GROUSE. Furthermore, our guarantees hold with high probability, rather than in expectation, an improvement made possible by the machinery deployed in this work. Most importantly, SNIPE addresses the lack of global convergence in GROUSE with missing data. Indeed, the first measurement block in SNIPE provides a good initialization (see Lemma 2) which, combined with Lemma 3, guarantees global convergence of SNIPE to the true subspace (see Theorem 1).

Lastly, both SNIPE and GROUSE have a computational complexity of $O(rn)$ floating-point operations per incoming vector (see Remark 6). Also, SNIPE for $k \geq 2$ and GROUSE both require $O(rn)$ memory elements of storage. In theory, SNIPE requires a large first block $b_1 = O(n)$ and consequently $O(n^2)$ memory elements of storage (see Remark 7). In all simulations in Section 5, however, we set $b_1 = O(r)$, requiring a storage of $O(rn)$ memory elements. With full measurements ($p = 1$), a close relative of both SNIPE and GROUSE is incremental SVD [20, 21, 22].

We must also discuss the algorithms in [10, 16, 17, 18]. The algorithm in [10], in a sense, extends the classic power method to handle missing data, in part by improving the main result of [23]. With high probability, this algorithm converges globally and linearly to the true subspace and, most notably, succeeds for arbitrarily small sampling probability p , if the scope of the algorithm T is large enough. Additionally, this algorithm too has a computational complexity of $O(rn)$ operations per vector and a storage requirement of $O(rn)$ memory elements.

PETRELS, introduced in [16], operates on one column at a time (rather than blocks) and convergence is known only to a stationary point of (14), with global convergence an apparently open question. Designed for online *matrix completion*, the algorithm in [17] also operates on one column at a time (rather than blocks) and asymptotic convergence to the true subspace is established (see Propositions 2 and 3). This framework is also extended to tensors. MOUSSE in [18] tracks a union of subspaces (manifold) rather than just one; SNIPE would function more like an ingredient of this algorithm. Asymptotic consistency of MOUSSE is also established there. We point out that the theoretical guarantees for SNIPE surpass those of other algorithms; in Section 3, we showed nonasymptotic global convergence of SNIPE.

In the next section, we compare the performance of three of these algorithms in practice and find that SNIPE displays state-of-the-art performance.

As discussed in Section 1, SNIPE is also useful in Big Data applications where data could only be processed in small amounts. In particular, SNIPE is closely related to matrix completion algorithms. For example, consider a rank- r matrix $M \in \mathbb{R}^{n \times T}$ with column span of $\mathcal{S} \in \mathbb{G}(n, r)$. Recovering M from its entries on a (typically small) index set $\Omega \subset [1 : n] \times [1 : T]$ is known as the matrix completion problem [24, 25]. If T is large, then one might complete $M \in \mathbb{R}^{n \times T}$ by partitioning it into blocks of size $n \times b$, processing each block separately and in sequence, and cycling through the blocks more than once, if desired. Note that, given $\mathcal{S} = \text{span}(M)$, completing M reduces to a least-squares problem. Therefore, each iteration of SNIPE might be adjusted to complete the corresponding block of M based on the current estimate of \mathcal{S} , cycling through the measurement blocks in sequence until convergence. Note, however, that our analysis in Theorem 1 requires statistical independence between consecutive blocks and does not apply in this form if measurement blocks are cycled more than once. We also refer the interested reader to the discussions in on this topic [13, 10].

5 Simulations

This section consists of two parts: first, we empirically study the dependence of SNIPE on various parameters, and second we compare SNIPE with GROUSE and the algorithm in [10]. In all simulations, we consider an r -dimensional subspace $\mathcal{S} \subset \mathbb{R}^n$ and a sequence of generic vectors $\{s_t\}_{t=1}^T \subset \mathcal{S}$. Every entry of these vectors is observed with probability $p \in (0, 1]$ and collected in measurement vectors $\{y_t\}_{t=1}^T \subset \mathbb{R}^n$. Our objective is to estimate \mathcal{S} from $\{y_t\}$, as described in Section 1.

Sampling probability We first set $n = 100$, $r = 5$, and let \mathcal{S} be a generic r -dimensional subspace, namely the span of an $n \times r$ standard random Gaussian matrix. For various values of probability p , we run SNIPE with block size $b = b_1 = b_2 = \dots = b_K = 2r = 10$ and scope of $T = 500r = 2500$, recording the average (over 50 trials) estimation error $d_{\mathbb{G}}(\mathcal{S}, \hat{\mathcal{S}}_K)$ (see (6)). The average error versus probability is plotted in Figure 3a.

Subspace dimension With the same setting as the previous paragraph, we now set $p = 3r/n = 0.15$ and vary the subspace dimension r , block size $b = 2r$, and scope $T = 500r$. The average error versus subspace dimension is plotted in Figure 3b.

Ambient dimension This time, we set $r = 5$, $p = 3r/n$, $b = 2r$, $T = 500r$, and vary the ambient dimension n . In other words, we vary n while keeping the number of measurements per vector fixed at $\sim pn = 3r$. The average error versus ambient dimension is plotted in Figure 3c. Observe that the performance of SNIPE steadily degrades as n increases. This is in agreement with Lemma 3 which, roughly speaking, sets the convergence rate at $\sqrt{1-p}$; as n increases, $p = 3r/n$ and the convergence rate both drop. See also Theorem 2.14 in [9].

Block size Next, we set $n = 100$, $r = 5$, $p = 3r/n$, $T = 500r$, and vary the block size $b = b_1 = b_2 = \dots = b_K$. The average error versus block size in both cases is depicted in Figure 3d. As discussed in Remark 5, a block size of $b = \Omega(r)$ is necessary for the success of SNIPE, which explains the poor performance of SNIPE for very small values of b . However, as b increases, the number of blocks $K = T/b$ reduces because the scope T is held fixed. As the estimation error of SNIPE scales like 2^{-K} in Theorem 1, the performance suffers in Figure 3d. It appears that the choice of $b = Cr$ in SNIPE for a relatively small C (like $C = 2$) guarantees the best performance.

Coherence Lastly, we set $n = 300$, $r = 10$, $p = 3r/n$, $b = 2r$, and $T = 500r$. We then test the performance of SNIPE as the coherence of \mathcal{S} varies (see (7)). To that end, let $\mathcal{S} \subset \mathbb{R}^n$ be a generic subspace with orthonormal basis $S \in \mathbb{R}^{n \times r}$. Then, the average coherence of \mathcal{S} over 50 trials was $3.3334 \ll n/r$ and the average estimation error of SNIPE was $2.795 \cdot 10^{-5}$. On the other hand, let $D \in \mathbb{R}^{n \times n}$ be a diagonal matrix with entries $D[i, i] = i^{-1}$ and consider $S' = DS$. Unlike \mathcal{S} , $\mathcal{S}' := \text{span}(S')$ is typically a coherent subspace since the energy of S' is mostly concentrated along its first few rows. This time, the average coherence of \mathcal{S}' over 50 trials was $19.1773 \approx n/r$ and the average estimation error of SNIPE was 0.4286.

Comparisons Next, we empirically compare SNIPE with GROUSE and the modified power method in [10]. We set $n = 100$, $r = 5$, $T = 5 \cdot 10^3$, and take $\mathcal{S} \subset \mathbb{R}^n$ to be a generic r -dimensional subspace. For $p = 3r/n = 0.15$, we compare the three algorithms in Figure 4a, which shows the average over 100 trials of the estimation error of algorithms (with respect to the metric $d_{\mathbb{G}}$) as time progresses. For SNIPE, we used the block size of $b = 2r$. Having tried to get the best performance from GROUSE, the (diminishing) step size is set to $100/t$. For [10], we set the block size as $b = 2n$ for the best outcome.

We also compare, for various values of the sampling probability p , the final estimation error of the three algorithms in Figure 4b. In both tests, SNIPE is comparable to GROUSE, and substantially improves over the power method. Note also that SNIPE, unlike many similar algorithms, has provable global convergence and, because it operates on measurement blocks, we suspect that it will show more robustness against noise. We leave that investigation to a future work.

6 Theory

In this section, we prove Theorem 1 in two steps. First, in Section 6.1, we establish that the first iteration of SNIPE provides a good initial estimate of the true subspace \mathcal{S} . Then, in Section 6.2, we analyse how SNIPE iteratively refines this initial estimate.

A short word on notation is in order first. We will frequently use MATLAB's matrix notation so that, for example, $A[i, j]$ is the $[i, j]$ th entry of A , and the row-vector $A[i, :]$ corresponds to the i th row of A . By

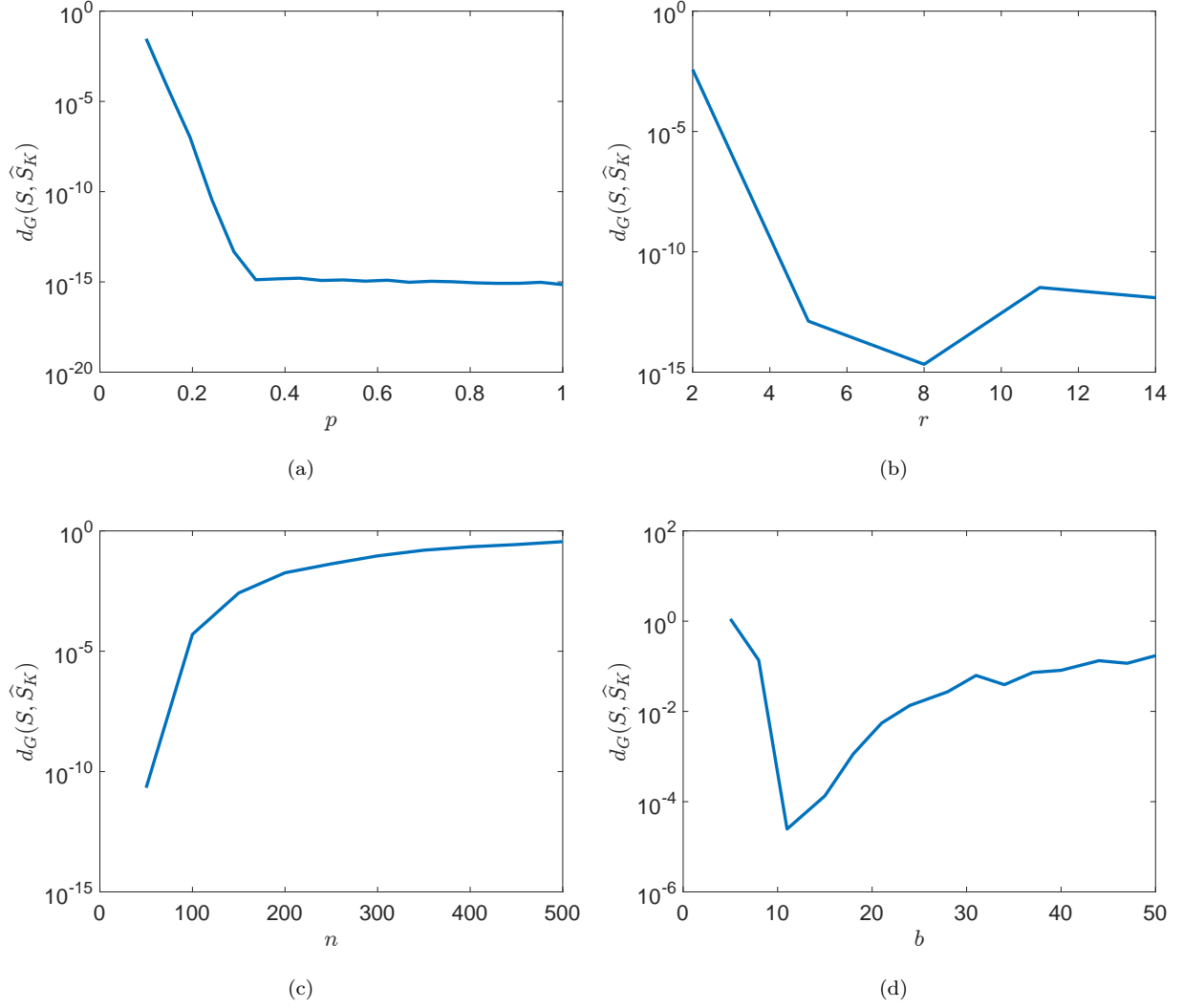


Figure 3: Performance of SNIPE as (a) sampling probability p , (b) subspace dimension r , (c) ambient dimension n , (d) block size b vary. $\hat{\mathcal{S}}_K$ is the output of SNIPE and $d_G(\mathcal{S}, \hat{\mathcal{S}}_K)$ is its distance to the true subspace \mathcal{S} , which generated the input of SNIPE. See Section 5 for details.

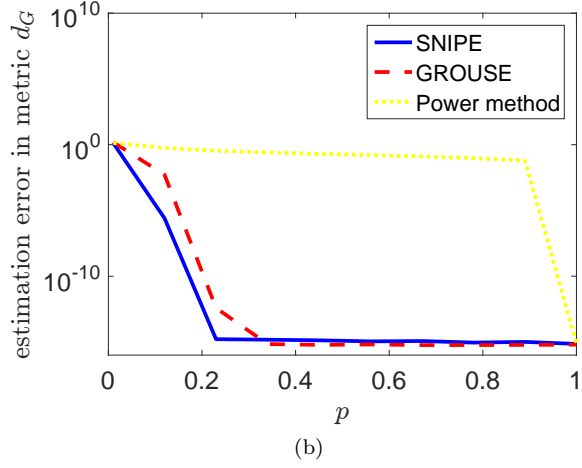
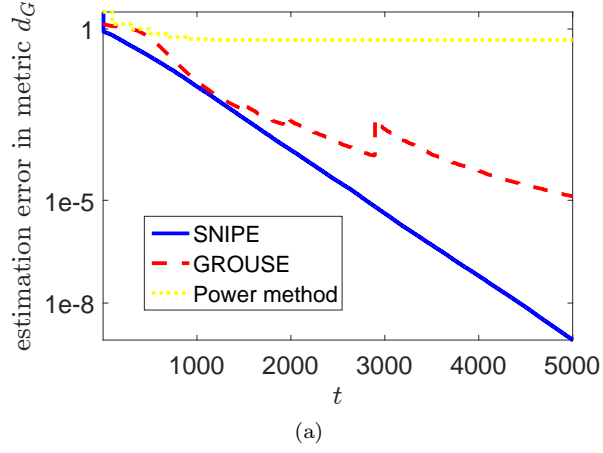


Figure 4: (a) Estimation error versus time for SNIPE, GROUSE, and the modified power method in [10] with a prescribed set of parameters. (b) Estimation error of these algorithms as the sampling probability p varies. See Section 5 for details.

$\{\epsilon_i\}_i \stackrel{\text{ind.}}{\sim} \text{Bernoulli}(p)$, we mean that $\{\epsilon_i\}_i$ are independent Bernoulli random variables taking one with probability of p and zero otherwise. Throughout, $E_{i,j}$ stands for the $[i,j]$ th canonical matrix so that $E_{i,j}[i,j] = 1$ is its only nonzero entry. The size of $E_{i,j}$ may be inferred from the context. As usual, $\|\cdot\|$ and $\|\cdot\|_F$ stand for the spectral and Frobenius norms. In addition, $\|A\|_\infty$ and $\|A\|_{2 \rightarrow \infty}$ return the largest entry of a matrix A (in magnitude) and the largest ℓ_2 norm of the rows of A , respectively.

For purely aesthetic reasons, we will occasionally use the notation $a \vee b = \max(a, b)$ and $a \wedge b = \min(a, b)$.

6.1 Initialization

In the first iteration ($k = 1$), SNIPE simply sets $\hat{\mathcal{S}}_1$ to be the span of the top r (left) singular vectors of the first incoming measurement block Y_1 . The following lemma, roughly speaking, bounds the largest principal angle between \mathcal{S} and $\hat{\mathcal{S}}_1$ as

$$\theta_1(\mathcal{S}, \hat{\mathcal{S}}_1) \lesssim \sqrt{\frac{r}{pn}},$$

and is proved in Appendix C with the aid of standard large deviation bounds.

Lemma 2. *For $\alpha \geq 1$, $\nu \geq 1$, and $1 \leq \eta_1 \leq \frac{b_1}{r}$, it holds that*

$$\frac{\|P_{\mathcal{S}^\perp} P_{\hat{\mathcal{S}}_1}\|_F}{\sqrt{r}} \lesssim \alpha \cdot \nu \sqrt{\left(1 \vee \frac{n}{b_1}\right) \frac{(\eta_1 \vee \eta(\mathcal{S})) r \log(n \vee b_1)}{pn}}, \quad (14)$$

except with a probability of at most $e^{-\alpha} + \Pr[\nu(Q_1) > \nu] + \Pr[\eta(Q_1) > \eta_1]$.

6.2 Refinement

At iteration $k \in [2 : K]$, SNIPE uses the current estimate $\hat{\mathcal{S}}_{k-1}$ and the new measurement block Y_k to produce a new estimate $\hat{\mathcal{S}}_k$ (of the true subspace \mathcal{S}). The main challenge here is to compare the principal angles $\theta_1(\mathcal{S}, \hat{\mathcal{S}}_k)$ and $\theta_1(\mathcal{S}, \hat{\mathcal{S}}_{k-1})$. The following result, proved in Appendix D, roughly states that

$$\theta_1(\mathcal{S}, \hat{\mathcal{S}}_k) \lesssim \sqrt{1-p} \cdot \theta_1(\mathcal{S}, \hat{\mathcal{S}}_{k-1}),$$

provided that $\theta_1(\mathcal{S}, \hat{\mathcal{S}}_{k-1}) \lesssim \sqrt{p}$.

Lemma 3. *Fix $k \in [2 : K]$, $\alpha \geq 1$, $\nu \geq 1$, and $1 \leq \eta_k \leq \frac{b}{r}$. Then, except with a probability of at most $O(e^{-\alpha} + b_k^{1-\alpha}) + \Pr[\nu(Q_k) > \nu] + \Pr[\eta(Q_k) > \eta_k]$, it holds that*

$$\frac{\|P_{\mathcal{S}^\perp} P_{\hat{\mathcal{S}}_k}\|_F}{\sqrt{r}} \lesssim \left(\alpha \cdot \nu \left(\sqrt{\frac{\eta_k}{b_k}} + \sqrt{1-p} \right) + \frac{1}{4} \right) \frac{\|P_{\mathcal{S}^\perp} P_{\hat{\mathcal{S}}_{k-1}}\|_F}{\sqrt{r}}, \quad (15)$$

provided that $p \gtrsim \alpha^2 \log^3(n \vee b_k) \cdot \eta(\hat{\mathcal{S}}_{k-1})r/n$, and

$$\|P_{\mathcal{S}^\perp} P_{\hat{\mathcal{S}}_{k-1}}\|_F \lesssim \frac{\sqrt{p}}{\alpha \cdot \nu \log^{\frac{3}{2}}(n \vee b_k)}. \quad (16)$$

6.3 Completing the Proof of Theorem 1

We now combine Lemmas 2 and 3 to conclude the analysis of SNIPE. Fix $\alpha \geq 1$, $\nu \geq 1$, and $\{\eta_k\}_{k=1}^K$ with $1 \leq \eta_k \leq b_k/r$ for every k . Conveniently, set $b_{\max} := \max_k b_k$. Recall that SNIPE produces the sequence of estimates $\{\hat{\mathcal{S}}_k\}_{k=1}^K$. In light of Lemma 2, for $\hat{\mathcal{S}}_1$ to be a good initialization that would “activate” (16) in Lemma 3 with $k = 2$, it suffices to have that

$$\alpha \cdot \nu \sqrt{\left(1 \vee \frac{n}{b_1}\right) \frac{(\eta_1 \vee \eta(\mathcal{S})) r \log(n \vee b_1)}{pn}} \lesssim \frac{1}{\alpha \cdot \nu \log^{\frac{3}{2}}(n \vee b_2)} \sqrt{\frac{p}{r}}. \quad (17)$$

That is, under (17), both (14) and (16) (with $k = 2$) hold except with a probability of at most $e^{-\alpha} + \Pr[\nu(Q_1) > \nu] + \Pr[\eta(\mathcal{Q}_1) > \eta_1]$. In other words, $\hat{\mathcal{S}}_1$ is a good initial estimate for the true subspace and, from this point onward, the estimation error reduces exponentially fast, as dictated by (15). To quantify this exponential convergence, next suppose that

$$\alpha \cdot \nu \left(\sqrt{\frac{\eta_k}{b_k}} + \sqrt{1-p} \right) \lesssim 1, \quad \forall k \in [1 : K-1], \quad (18)$$

so that, in particular, the factor in front of $\|P_{\mathcal{S}^\perp} P_{\hat{\mathcal{S}}_{k-1}}\|_F / \sqrt{r}$ in (15) does not exceed $1/2$. In particular, as long as $\eta_k \lesssim b_k / (\alpha \cdot \nu)^2$, (18) simplifies to

$$p \geq 1 - \frac{C_1}{\alpha^2 \nu^2}, \quad (19)$$

for an absolute constant $C_1 > 0$. In light of (19), (17) holds when

$$\alpha \cdot \nu \sqrt{\left(1 \vee \frac{n}{b_1}\right) \frac{(\eta_1 \vee \eta(\mathcal{S})) r \log(n \vee b_1)}{(1 - \frac{C_1}{\alpha^2 \nu^2}) n}} \lesssim \frac{1}{\alpha \cdot \nu \log^{\frac{3}{2}}(n \vee b_2)} \sqrt{\frac{p}{r}}, \quad (20)$$

or

$$p \gtrsim \frac{\alpha^6 \nu^6}{\alpha^2 \nu^2 - C_1} \left(1 \vee \frac{n}{b_1}\right) \frac{(\eta_1 \vee \eta(\mathcal{S})) r^2 \log^4(n \vee b_{\max})}{n}. \quad (21)$$

Assume, without loss of generality, that $\|P_{\mathcal{S}^\perp} P_{\hat{\mathcal{S}}_k}\|_F \neq 0$ for every $k \in [1 : K-1]$. Otherwise, by Lemma 3, SNIPE converges to \mathcal{S} early, with high probability. Then, for every $k \in [2 : K]$, Lemmas 2 and 3 imply that

$$\begin{aligned} & \frac{\|P_{\mathcal{S}^\perp} P_{\hat{\mathcal{S}}_k}\|_F}{\sqrt{r}} \\ &= \prod_{k'=2}^k \frac{\|P_{\mathcal{S}^\perp} P_{\hat{\mathcal{S}}_{k'}}\|_F}{\|P_{\mathcal{S}^\perp} P_{\hat{\mathcal{S}}_{k'-1}}\|_F} \cdot \frac{\|P_{\mathcal{S}^\perp} P_{\hat{\mathcal{S}}_1}\|_F}{\sqrt{r}} \\ &\lesssim \prod_{k'=2}^k \left(\alpha \cdot \nu \left(\sqrt{\frac{\eta_{k'}}{b_{k'}}} + \sqrt{1-p} \right) + \frac{1}{4} \right) \cdot \alpha \cdot \nu \sqrt{\left(1 \vee \frac{n}{b_1}\right) \frac{(\eta_1 \vee \eta(\mathcal{S})) r \log(n \vee b_{\max})}{pn}} \quad (\text{Lemmas 2 and 3}) \\ &\lesssim 2^{-k+1} \cdot \alpha \cdot \nu \sqrt{\left(1 \vee \frac{n}{b_1}\right) \frac{(\eta_1 \vee \eta(\mathcal{S})) r \log(n \vee b_{\max})}{pn}}, \quad (\text{see (18)}) \end{aligned} \quad (22)$$

provided that

$$p \gtrsim \alpha^2 \max_{k \in [1:K-1]} \eta(\hat{\mathcal{S}}_k) \cdot \frac{r \log^3(n \vee b_{\max})}{n}, \quad (23)$$

and except with a probability of at most

$$K \cdot O \left(e^{-\alpha} + (\min_{k'} b_{k'})^{1-\alpha} \right) + \sum_{k'=1}^K \Pr[\nu(Q_{k'}) > \nu] + \Pr[\eta(\mathcal{Q}_{k'}) > \eta_{k'}]. \quad (24)$$

In words, (22) states that the estimation error is halved in every iteration, with high probability and provided that (19), (21), and (23) hold.

In particular, to better interpret (23), we must replace the coherence $\eta(\hat{\mathcal{S}}_k)$ therein with a simpler quantity, perhaps $\eta(\mathcal{S})$. We can do so thanks to Lemma 7 in the Toolbox (Appendix A) which, roughly speaking, states that a pair of subspaces \mathcal{A} and \mathcal{B} with a small principal angle have similar coherences, namely $\theta_1(\mathcal{A}, \mathcal{B}) \approx 0 \implies \eta(\mathcal{A}) \approx \eta(\mathcal{B})$. More concretely, using (22) and after invoking Lemma 7, we find the

following for every $k \in [1 : K]$:

$$\begin{aligned}
\sqrt{\eta(\widehat{\mathcal{S}}_k)} &\leq \sqrt{\eta(\mathcal{S})} + \|P_{\mathcal{S}^\perp} P_{\mathcal{S}_k}\| \sqrt{\frac{n}{r}} \quad (\text{see Lemma 7}) \\
&\leq \sqrt{\eta(\mathcal{S})} + \|P_{\mathcal{S}^\perp} P_{\mathcal{S}_k}\|_F \sqrt{\frac{n}{r}} \\
&\lesssim \sqrt{\eta(\mathcal{S})} + \sqrt{r} \left(2^{-k+1} \cdot \alpha \cdot \nu \sqrt{\left(1 \vee \frac{n}{b_1}\right) \frac{(\eta_1 \vee \eta(\mathcal{S})) r \log(n \vee b_{\max})}{pn}} \right) \sqrt{\frac{n}{r}} \quad (\text{see (22)}) \\
&= \sqrt{\eta(\mathcal{S})} + 2^{-k+1} \cdot \alpha \cdot \nu \sqrt{\left(1 \vee \frac{n}{b_1}\right) \frac{(\eta_1 \vee \eta(\mathcal{S})) r \log(n \vee b_{\max})}{p}} \\
&\leq \sqrt{\eta(\mathcal{S})} + \alpha \cdot \nu \sqrt{\left(1 \vee \frac{n}{b_1}\right) \frac{(\eta_1 \vee \eta(\mathcal{S})) r \log(n \vee b_{\max})}{p}} \\
&\lesssim \alpha \cdot \nu \sqrt{\left(1 \vee \frac{n}{b_1}\right) \frac{(\eta_1 \vee \eta(\mathcal{S})) r \log(n \vee b_{\max})}{p}} \\
&\leq \alpha \cdot \nu \sqrt{\left(1 \vee \frac{n}{b_1}\right) \frac{(\eta_1 \vee \eta(\mathcal{S})) r \log(n \vee b_{\max})}{1 - C_1/\alpha^2 \nu^2}}. \quad (\text{see (19)})
\end{aligned}$$

Consequently, (23) holds when

$$p \gtrsim \frac{\alpha^6 \nu^4}{\alpha^2 \nu^2 - C_1} \left(1 \vee \frac{n}{b_1}\right) \frac{(\eta_1 \vee \eta(\mathcal{S})) r^2 \log^4(n \vee b_{\max})}{n}. \quad (25)$$

To summarize, the exponential convergence promised in (22) is valid if (19) and (21) hold (because (21) implies (25) too). This completes the proof of Theorem 1.

Acknowledgements

AE would like to thank Anand Vidyashankar and Chris Williams for separately pointing out the possibility of a statistical interpretation of SNIPE, as discussed at the end of Section 2.1. MBW was partially supported by NSF grant CCF-1409258 and NSF CAREER grant CCF-1149225.

References

- [1] P. van Overschee and B. L. de Moor. *Subspace identification for linear systems: Theory, implementation, applications*. Springer US, 2012.
- [2] B. A. Ardekani, J. Kershaw, K. Kashikura, and I. Kanno. Activation detection in functional MRI using subspace modeling and maximum likelihood estimation. *IEEE Transactions on Medical Imaging*, 18(2):101–114, 1999.
- [3] H. Krim and M. Viberg. Two decades of array signal processing research: The parametric approach. *IEEE Signal processing magazine*, 13(4):67–94, 1996.
- [4] L. Tong and S. Perreau. Multichannel blind identification: From subspace to maximum likelihood methods. *Proceedings of IEEE*, 86:1951–1968, 1998.
- [5] J. P. Cunningham and Z. Ghahramani. Linear dimensionality reduction: Survey, insights, and generalizations. *Journal of Machine Learning Research*, 16:2859–2900, 2015.
- [6] B. Recht, C. Re, S. J. Wright, and F. Niu. Hogwild: A lock-free approach to parallelizing stochastic gradient descent. In *Advances in Neural Information Processing Systems*, pages 693–701, 2011.

- [7] J. Nocedal and S. J. Wright. *Numerical optimization*. Springer Series in Operations Research and Financial Engineering. Springer New York, 2006.
- [8] G. H. Golub and C. F. Van Loan. *Matrix computations*. Johns Hopkins Studies in the Mathematical Sciences. Johns Hopkins University Press, 2013.
- [9] L. Balzano and S. J. Wright. Local convergence of an algorithm for subspace identification from partial data. *Foundations of Computational Mathematics*, 15(5):1279–1314, 2015.
- [10] I. Mitliagkas, C. Caramanis, and P. Jain. Streaming PCA with many missing entries. *Preprint*, 2014.
- [11] Y. Chen. Incoherence-optimal matrix completion. *IEEE Transactions on Information Theory*, 61(5):2909–2923, 2015.
- [12] F. W. J. Olver, National Institute of Standards, and Technology (U.S.). *NIST Handbook of Mathematical Functions Hardback and CD-ROM*. Cambridge University Press, 2010.
- [13] L. Balzano, R. Nowak, and B. Recht. Online identification and tracking of subspaces from highly incomplete information. In *Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pages 704–711. IEEE, 2010.
- [14] N. Halko and J. A. Martinsson, P. G. and Tropp. Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions. *SIAM review*, 53(2):217–288, 2011.
- [15] R. Ge, J. D. Lee, and T. Ma. Matrix completion has no spurious local minimum. *arXiv preprint arXiv:1605.07272*, 2016.
- [16] Y. Chi, Y. C. Eldar, and R. Calderbank. PETRELS: Parallel subspace estimation and tracking by recursive least squares from partial observations. *IEEE Transactions on Signal Processing*, 61(23):5947–5959, 2013.
- [17] M. Mardani, G. Mateos, and G. B. Giannakis. Subspace learning and imputation for streaming big data matrices and tensors. *IEEE Transactions on Signal Processing*, 63(10):2663–2677, 2015.
- [18] Y. Xie, J. Huang, and R. Willett. Change-point detection for high-dimensional time series with missing data. *IEEE Journal of Selected Topics in Signal Processing*, 7(1):12–27, 2013.
- [19] L. Balzano and S. J. Wright. On GROUSE and incremental SVD. In *IEEE International Workshop on Computational Advances in Multi-Sensor Adaptive Processing (CAMSAP)*, pages 1–4. IEEE, 2013.
- [20] J. R. Bunch and C. P. Nielsen. Updating the singular value decomposition. *Numerische Mathematik*, 31(2):111–129, 1978.
- [21] A. Balsubramani, S. Dasgupta, and Y. Freund. The fast convergence of incremental pca. In *Advances in Neural Information Processing Systems*, pages 3174–3182, 2013.
- [22] E. Oja and J. Karhunen. On stochastic approximation of the eigenvectors and eigenvalues of the expectation of a random matrix. *Journal of mathematical analysis and applications*, 106(1):69–84, 1985.
- [23] K. Lounici. High-dimensional covariance matrix estimation with missing observations. *Bernoulli*, 20(3):1029–1058, 2014.
- [24] E. J. Candès and B. Recht. Exact matrix completion via convex optimization. *Foundations of Computational mathematics*, 9(6):717–772, 2009.
- [25] A. Eftekhari, M. B. Wakin, and R. A. Ward. Mc2: A two-phase algorithm for leveraged matrix completion. *arXiv preprint arXiv:1609.01795*, 2016.
- [26] D. Gross. Recovering low-rank matrices from few coefficients in any basis. *IEEE Transactions on Information Theory*, 57(3):1548–1566, 2011.

- [27] J. A. Tropp. User-friendly tail bounds for sums of random matrices. *Foundations of Computational Mathematics*, 12(4):389–434, 2012.
- [28] P. Wedin. Perturbation bounds in connection with singular value decomposition. *BIT Numerical Mathematics*, 12(1):99–111, 1972.
- [29] R. Vershynin. Introduction to the non-asymptotic analysis of random matrices. In Y. C. Eldar and G. Kutyniok, editors, *Compressed Sensing: Theory and Applications*, pages 95–110. Cambridge University Press, 2012.

A Toolbox

This section collects a number of useful results for the reader’s convenience. We begin with the following concentration inequalities that are repeatedly invoked in the rest of the appendices [26, 27].

Lemma 4. [Matrix Bernstein inequality for Frobenius norm] *Let $\{Z_i\} \subset \mathbb{R}^{n \times b}$ be a finite sequence of zero-mean independent random matrices, and set*

$$\beta := \max_i \|Z_i\|_F,$$

$$\sigma^2 := \sum_i \mathbb{E} \|Z_i\|_F^2.$$

Then, for $\alpha \geq 1$ and except with a probability of at most $e^{-\alpha}$, it holds that

$$\left\| \sum_i Z_i \right\|_F \lesssim \alpha \cdot \max(\beta, \sigma).$$

Lemma 5. [Matrix Bernstein inequality for spectral norm] *Let $\{Z_i\} \subset \mathbb{R}^{n \times b}$ be a finite sequence of zero-mean independent random matrices, and set*

$$\beta := \max_i \|Z_i\|,$$

$$\sigma^2 := \left\| \sum_i \mathbb{E} [Z_i^* Z_i] \right\| \vee \left\| \sum_i \mathbb{E} [Z_i Z_i^*] \right\|.$$

Then, for $\alpha \geq 1$ and except with a probability of at most $e^{-\alpha}$, it holds that

$$\left\| \sum_i Z_i \right\| \lesssim \alpha \cdot \max \left(\log(n \vee b) \cdot \beta, \sqrt{\log(n \vee b)} \cdot \sigma \right).$$

Consider r -dimensional subspaces \mathcal{A} and \mathcal{B} . A review of the following well-known identities is perhaps helpful:

$$\sin(\theta_1(\mathcal{A}, \mathcal{B})) = \|P_{\mathcal{A}^\perp} P_{\mathcal{B}}\| = \|P_{\mathcal{A}} - P_{\mathcal{B}}\|, \quad (26)$$

$$\sqrt{\sum_{i=1}^r \sin^2(\theta_i(\mathcal{A}, \mathcal{B}))} = \|P_{\mathcal{A}^\perp} P_{\mathcal{B}}\|_F = \frac{1}{\sqrt{2}} \|P_{\mathcal{A}} - P_{\mathcal{B}}\|_F. \quad (27)$$

The following lemma is a simple variation of the standard perturbation bounds [28].

Lemma 6. [Perturbation bound] Fix a rank- r matrix A and let $\mathcal{A} = \text{span}(A)$. For matrix B , let B_r be a rank- r truncation of B obtained via SVD, and set $\mathcal{B}_r = \text{span}(B_r)$. Then, it holds that

$$\begin{aligned} \|P_{\mathcal{A}} - P_{\mathcal{B}_r}\| &= \|P_{\mathcal{A}^\perp} P_{\mathcal{B}_r}\| \leq \frac{\|P_{\mathcal{A}^\perp} B\|}{\sigma_r(A) - \|B - A\|} \leq \frac{\|B - A\|}{\sigma_r(A) - \|B - A\|}, \\ \frac{1}{\sqrt{2}} \|P_{\mathcal{A}} - P_{\mathcal{B}_r}\|_F &= \|P_{\mathcal{A}^\perp} P_{\mathcal{B}_r}\|_F \leq \frac{\|P_{\mathcal{A}^\perp} B\|_F}{\sigma_r(A) - \|B - A\|} \leq \frac{\|B - A\|_F}{\sigma_r(A) - \|B - A\|}, \end{aligned}$$

where $\sigma_r(A)$ is the smallest nonzero singular value of A .

Proof. Let $B_{r+} := B - B_r$ denote the residual and note that

$$\begin{aligned} \|P_{\mathcal{A}^\perp} P_{\mathcal{B}_r}\| &= \|P_{\mathcal{A}^\perp} B_r B_r^\dagger\| \quad (\mathcal{B}_r = \text{span}(B_r)) \\ &= \|P_{\mathcal{A}^\perp} (B - B_{r+}) B_r^\dagger\| \quad (B = B_r + B_{r+}) \\ &= \|P_{\mathcal{A}^\perp} B B_r^\dagger\| \quad (\text{span}(B_{r+}^\perp) \perp \text{span}(B_r^*) = \text{span}(B_r^\dagger)) \\ &\leq \|P_{\mathcal{A}^\perp} B\| \cdot \|B_r^\dagger\| \\ &= \frac{\|P_{\mathcal{A}^\perp} B\|}{\sigma_r(B_r)} \\ &\leq \frac{\|P_{\mathcal{A}^\perp} B\|}{\sigma_r(A) - \|B - A\|}. \quad (\text{Weyl's inequality}) \end{aligned}$$

The proof is identical for the claim with the Frobenius norm and is omitted here. \square

Lemma 7. [Coherence under perturbation] Let \mathcal{A}, \mathcal{B} be two r -dimensional subspaces in \mathbb{R}^n , and let $\theta_1(\mathcal{A}, \mathcal{B})$ be their (largest) principal angle. Then, their coherences are related as follows:

$$\sqrt{\eta(\mathcal{B})} \leq \sqrt{\eta(\mathcal{A})} + \|P_{\mathcal{A}^\perp} P_{\mathcal{B}}\| \sqrt{\frac{n}{r}}.$$

Proof. Let $\theta_1 \geq \theta_2 \geq \dots \geq \theta_r$ be the principal angles between the subspaces \mathcal{A} and \mathcal{B} so that, in particular, $\theta_1 = \theta_1(\mathcal{A}, \mathcal{B})$. It is well-known [8] that there exist orthonormal bases $A, B \in \mathbb{R}^{n \times r}$ for the subspaces \mathcal{A} and \mathcal{B} , respectively, such that

$$A^* B = \text{diag} \left(\begin{bmatrix} \cos \theta_1 & \cos \theta_2 & \dots & \cos \theta_r \end{bmatrix} \right) =: \Gamma \in \mathbb{R}^{r \times r}, \quad (28)$$

where $\text{diag}(a)$ is the diagonal matrix formed from vector a . There also exists $A' \in \mathbb{R}^{n \times r}$ with orthonormal columns such that

$$(A')^* B = \text{diag} \left(\begin{bmatrix} \sin \theta_1 & \sin \theta_2 & \dots & \sin \theta_r \end{bmatrix} \right) =: \Sigma \in \mathbb{R}^{r \times r}, \quad (A')^* A = 0, \quad (29)$$

and, moreover,

$$\text{span} \left(\begin{bmatrix} A & B \end{bmatrix} \right) = \text{span} \left(\begin{bmatrix} A & A' \end{bmatrix} \right). \quad (30)$$

With $\mathcal{A}' = \text{span}(A')$, it follows that

$$\begin{aligned} B &= P_{\mathcal{A}} B + P_{\mathcal{A}'} B \\ &= A A^* B + A' (A')^* B \\ &= A \Gamma + A' \Sigma. \quad (\text{see (28) and (29)}) \end{aligned} \quad (31)$$

Consequently,

$$\begin{aligned}
\sqrt{\eta(\mathcal{B})} &= \sqrt{\frac{n}{r} \max_i \|B[i, :]\|_2} \quad (\text{see (7)}) \\
&\leq \sqrt{\frac{n}{r} \max_i \|A[i, :] \cdot \Gamma\|_2} + \sqrt{\frac{n}{r} \max_i \|A'[i, :] \cdot \Sigma\|_2} \quad ((31) \text{ and triangle inequality}) \\
&\leq \sqrt{\frac{n}{r} \max_i \|A[i, :]\|_2 \|\Gamma\|} + \sqrt{\frac{n}{r} \max_i \|A'[i, :]\|_2 \|\Sigma\|} \\
&= \sqrt{\eta(\mathcal{A})} \|\Gamma\| + \sqrt{\eta(\mathcal{A}')} \|\Sigma\| \quad (\text{see (7)}) \\
&\leq \sqrt{\eta(\mathcal{A})} \|\Gamma\| + \sqrt{\frac{n}{r}} \|\Sigma\| \quad \left(\eta(\mathcal{A}') \leq \frac{n}{r}\right) \\
&\leq \sqrt{\eta(\mathcal{A})} + \sqrt{\frac{n}{r}} \sin \theta_1 \quad (\text{see (28) and (29)}) \\
&= \sqrt{\eta(\mathcal{A})} + \sqrt{\frac{n}{r}} \|P_{\mathcal{A}^\perp} P_{\mathcal{B}}\| \quad (\text{see (26)})
\end{aligned} \tag{32}$$

which completes the proof of Lemma 7. \square

B Supplement to Section 2.1

In this section, we verify that R_k (see (1)) is indeed the solution of

$$\begin{cases} \min \left\| P_{\widehat{S}_{k-1}^\perp} X_k \right\|_F^2, \\ P_{\Omega_k}(X_k) = Y_k, \end{cases} \tag{33}$$

when $k \geq 2$. First note that Program (33) is separable and equivalent to the following b programs:

$$R_k[:, j] = \arg \begin{cases} \min \left\| P_{\widehat{S}_{k-1}^\perp} x \right\|_2^2, \\ P_{\omega_t} \cdot x = y_t, \end{cases} \quad t = (k-1)b + j, \quad j \in [1 : b]. \tag{34}$$

Here, $R_k[:, j] \in \mathbb{R}^n$ is the j th column of R_k (in MATLAB's matrix notation). To solve each of the programs in (34), we make the change of variables $x = y_t + Q_{\omega_t^C} \cdot z$. Here, $z \in \mathbb{R}^{n-m}$ and $Q_{\omega_t^C} \in \{0, 1\}^{n \times (n-m)}$ is defined naturally so that $P_{\omega_t^C} = Q_{\omega_t^C} Q_{\omega_t^C}^*$. We now rewrite (34) as

$$z_j := \arg \max \left\| (S_{k-1}^\perp)^* y_t + (S_{k-1}^\perp)^* Q_{\omega_t^C} \cdot z \right\|_2^2, \quad t = (k-1)b + j, \quad j \in [1 : b]. \tag{35}$$

The solutions of these least-square programs are

$$z_j = - \left((S_{k-1}^\perp)^* Q_{\omega_t^C} \right)^\dagger (S_{k-1}^\perp)^* y_t, \quad j \in [1 : b]. \tag{36}$$

For fixed j , we simplify the expression for z_j as follows:

$$\begin{aligned}
z_j &= - \left((S_{k-1}^\perp)^* Q_{\omega_t^C} \right)^\dagger (S_{k-1}^\perp)^* y_t \\
&= - \left(Q_{\omega_t^C}^* P_{S_{k-1}^\perp} Q_{\omega_t^C} \right)^{-1} Q_{\omega_t^C}^* P_{S_{k-1}^\perp} y_t \quad (\text{almost surely, if } m \geq r??) \\
&= - \left(I_{n-m} - Q_{\omega_t^C}^* P_{S_{k-1}} Q_{\omega_t^C} \right)^{-1} Q_{\omega_t^C}^* P_{S_{k-1}^\perp} y_t \\
&= - \left(I_{n-m} + Q_{\omega_t^C}^* S_{k-1} \left(I_r - S_{k-1}^* P_{\omega_t^C} S_{k-1} \right)^{-1} S_{k-1}^* Q_{\omega_t^C} \right) Q_{\omega_t^C}^* P_{S_{k-1}^\perp} y_t \quad (\text{inversion lemma}) \\
&= - \left(I_{n-m} + Q_{\omega_t^C}^* S_{k-1} \left(I_r - S_{k-1}^* P_{\omega_t^C} S_{k-1} \right)^{-1} S_{k-1}^* Q_{\omega_t^C} \right) Q_{\omega_t^C}^* P_{S_{k-1}^\perp} P_{\omega_t} y_t \quad (y_t = P_{\omega_t} y_t) \\
&= \left(I_{n-m} + Q_{\omega_t^C}^* S_{k-1} \left(I_r - S_{k-1}^* P_{\omega_t^C} S_{k-1} \right)^{-1} S_{k-1}^* Q_{\omega_t^C} \right) Q_{\omega_t^C}^* P_{S_{k-1}} y_t \quad (Q_{\omega_t^C}^* Q_{\omega_t} = 0) \\
&= Q_{\omega_t^C}^* P_{S_{k-1}} y_t + Q_{\omega_t^C}^* S_{k-1} \left(I_r - S_{k-1}^* P_{\omega_t^C} S_{k-1} \right)^{-1} S_{k-1}^* P_{\omega_t^C} P_{S_{k-1}} y_t \\
&= Q_{\omega_t^C}^* P_{S_{k-1}} y_t + Q_{\omega_t^C}^* S_{k-1} \left(I_r - S_{k-1}^* P_{\omega_t^C} S_{k-1} \right)^{-1} \left(S_{k-1}^* P_{\omega_t^C} S_{k-1} - I_r \right) S_{k-1}^* y_t \\
&\quad + Q_{\omega_t^C}^* S_{k-1} \left(I_r - S_{k-1}^* P_{\omega_t^C} S_{k-1} \right)^{-1} S_{k-1}^* y_t \\
&= Q_{\omega_t^C}^* S_{k-1} \left(I_r - S_{k-1}^* P_{\omega_t^C} S_{k-1} \right)^{-1} S_{k-1}^* y_t \\
&= Q_{\omega_t^C}^* S_{k-1} \left(S_{k-1}^* P_{\omega_t} S_{k-1} \right)^{-1} S_{k-1}^* y_t \\
&= Q_{\omega_t^C}^* S_{k-1} (P_{\omega_t} S_{k-1})^\dagger y_t, \quad (y_t = P_{\omega_t} y_t)
\end{aligned} \tag{37}$$

which means that

$$y_t + Q_{\omega_t^C} z_j = y_t + P_{\omega_t^C} S_{k-1} (P_{\omega_t} S_{k-1})^\dagger y_t, \tag{38}$$

is the solution of the j th program in (34) which matches the j th column of R_k defined in (1).

C Proof of Lemma 2

Consider the coefficient vectors $\{q_t\}_{t=1}^{b_1} \subset \mathbb{R}^r$ and $\{s_t = Sq_t\}_{t=1}^{b_1} \subset \mathbb{R}^n$, constructed in Section 1. By concatenating $\{q_t\}$ and $\{s_t\}$, we form $Q_1 \in \mathbb{R}^{b_1 \times r}$ and $S_1 = SQ_1^* \in \mathbb{R}^{n \times b_1}$, respectively. Each s_t is observed on the random index set $\omega_t \subseteq [1 : n]$ or, equivalently, S_1 is observed on the random index set $\Omega_1 \subseteq [1 : n] \times [1 : b_1]$. We write this measurement process as $Y_1 = P_{\Omega_1}(S_1)$, where $P_{\Omega_1}(\cdot)$ projects onto index set Ω_1 .

Let us fix Q_1 for now. Also let $Y_{1,r} \in \mathbb{R}^{n \times b_1}$ be a rank- r truncation of Y_1 , obtained via SVD. SNIPE then sets $\hat{S}_1 = \text{span}(Y_{1,r})$. Our objective here is to control $\|P_{S^\perp} P_{\hat{S}_1}\|_F$. Since

$$\|P_{S^\perp} P_{\hat{S}_1}\|_F \leq \sqrt{r} \|P_{S^\perp} P_{\hat{S}_1}\|, \quad (\hat{S}_1 \in \mathbb{G}(n, r)) \tag{39}$$

it suffices to bound the spectral norm (instead of the Frobenius norm). Conditioned on Q_1 , it is easy to verify that $\mathbb{E}[Y_1] = \mathbb{E}[P_{\Omega_1}(S_1)] = p \cdot S_1$, suggesting that we might consider Y_1 as a perturbed copy of $p \cdot S_1$, and perhaps consider $\hat{S}_1 = \text{span}(Y_{1,r})$ as a perturbation of $\mathcal{S} = \text{span}(p \cdot S_1)$. Indeed, Lemma 6 in Toolbox (Appendix A) dictates that

$$\begin{aligned}
\|P_{S^\perp} P_{\hat{S}_1}\| &\leq \frac{\|Y_1 - pS_1\|}{p \cdot \sigma_r(S_1) - \|Y_1 - pS_1\|} \\
&= \frac{\|Y_1 - pS_1\|}{p \cdot \sigma_r(Q_1) - \|Y_1 - pS_1\|} \quad (S_1 = SQ_1^*, S^*S = I_r) \\
&\leq \frac{2}{p} \cdot \frac{\|Y_1 - pS_1\|}{\sigma_r(Q_1)}. \quad \left(\text{if } \|Y_1 - pS_1\| \leq \frac{p}{2} \cdot \sigma_r(Q_1) \right)
\end{aligned} \tag{40}$$

It remains to bound the norm in the last line above. To that end, we study the concentration of Y_1 about its expectation by writing that

$$\begin{aligned}
Y_1 - pS_1 &= P_{\Omega_1}(S_1) - pS_1 \\
&= \sum_{i,j} (\epsilon_{i,j} - p) S_1[i, j] \cdot E_{i,j} \\
&=: \sum_{i,j} Z_{i,j},
\end{aligned} \tag{41}$$

where $\{\epsilon_{i,j}\} \stackrel{\text{ind.}}{\sim} \text{Bernoulli}(p)$, and $E_{i,j} \in \mathbb{R}^{n \times b_1}$ is the $[i, j]$ th canonical matrix. Additionally, $\{Z_{i,j}\}$ are independent zero-mean random matrices. In order to appeal to the matrix Bernstein inequality (Lemma 5), we compute the β and σ parameters below, starting with β :

$$\begin{aligned}
\|Z_{i,j}\| &= \|(\epsilon_{i,j} - p) S_1[i, j] \cdot E_{i,j}\| \\
&= |(\epsilon_{i,j} - p) S_1[i, j]| \quad (\|E_{i,j}\| = 1) \\
&\leq |S_1[i, j]| \quad (\epsilon_{i,j} \in \{0, 1\}) \\
&\leq \|S_1\|_\infty \\
&\leq \|S\|_{2 \rightarrow \infty} \|Q_1\|_{2 \rightarrow \infty} \quad (S_1 = SQ_1^*, \|AB^*\|_\infty \leq \|A\|_{2 \rightarrow \infty} \|B\|_{2 \rightarrow \infty}) \\
&\leq \sqrt{\frac{\eta(\mathcal{S})r}{n}} \cdot \sqrt{\frac{\eta(Q_1)r}{b_1}} \cdot \|Q_1\| \quad (\text{see (7)}) \\
&=: \beta.
\end{aligned} \tag{42}$$

Above, $\|A\|_\infty$ and $\|A\|_{2 \rightarrow \infty}$, respectively, return the largest entry of A (in magnitude) and the largest ℓ_2 norm of the rows of matrix A . As for σ , we write that

$$\begin{aligned}
\left\| \mathbb{E} \left[\sum_{i,j} Z_{i,j} Z_{i,j}^* \right] \right\| &= \left\| \sum_{i,j} \mathbb{E} [(\epsilon_{i,j} - p)^2] S_1[i, j]^2 \cdot E_{i,i} \right\| \\
&= \left\| \sum_{i,j} p(1-p) S_1[i, j]^2 \cdot E_{i,i} \right\| \quad (\epsilon_{i,j} \sim \text{Bernoulli}(p)) \\
&\leq p \left\| \sum_{i,j} S_1[i, j]^2 \cdot E_{i,j} \right\| \\
&= p \left\| \sum_i \|S_1[i, :]\|_2^2 \cdot E_{i,i} \right\| \\
&= p \max_i \|S_1[i, :]\|_2^2 \\
&= p \|S_1\|_{2 \rightarrow \infty}^2 \\
&\leq p \|S\|_{2 \rightarrow \infty}^2 \cdot \|Q_1\|^2 \quad (S_1 = SQ_1^*, \|AB\|_{2 \rightarrow \infty} \leq \|A\|_{2 \rightarrow \infty} \|B\|) \\
&= p \cdot \frac{\eta(\mathcal{S})r}{n} \cdot \|Q_1\|^2. \quad (\text{see (7)})
\end{aligned} \tag{43}$$

In a similar fashion, we find that

$$\begin{aligned}
\left\| \mathbb{E} \left[\sum_{i,j} Z_{i,j}^* Z_{i,j} \right] \right\| &\leq p \left\| \sum_j \|S_1[:, j]\|_2^2 E_{j,j} \right\| \\
&= p \|S_1^*\|_{2 \rightarrow \infty}^2 \\
&\leq p \cdot \|S\|^2 \cdot \|Q_1\|_{2 \rightarrow \infty}^2 \quad (S_1 = SQ_1^*, \|AB\|_{2 \rightarrow \infty} \leq \|A\|_{2 \rightarrow \infty} \|B\|) \\
&\leq p \cdot \frac{\eta(Q_1) r}{b_1} \cdot \|Q_1\|^2, \quad (\|S\| = 1, \text{ see (7)})
\end{aligned} \tag{44}$$

and, eventually,

$$\begin{aligned}
\sigma^2 &= \left\| \mathbb{E} \left[\sum_{i,j} Z_{i,j}^* Z_{i,j} \right] \right\| \vee \left\| \mathbb{E} \left[\sum_{i,j} Z_{i,j} Z_{i,j}^* \right] \right\| \\
&\leq \frac{pr}{n} \left(1 \vee \frac{n}{b_1} \right) (\eta(\mathcal{S}) \vee \eta(Q_1)) \|Q_1\|^2. \quad (\text{see (43) and (44)})
\end{aligned} \tag{45}$$

Lastly,

$$\begin{aligned}
&\max \left(\log(n \vee b_1) \cdot \beta, \sqrt{\log(n \vee b_1)} \cdot \sigma \right) \\
&\lesssim \max \left(\log(n \vee b_1) \cdot \frac{r}{n}, \sqrt{\log(n \vee b_1)} \cdot \sqrt{\frac{pr}{n}} \right) \sqrt{1 \vee \frac{n}{b_1}} \cdot \sqrt{\eta(\mathcal{S}) \vee \eta(Q_1)} \cdot \|Q_1\| \quad (\text{see (42) and (45)}) \\
&\leq \sqrt{\log(n \vee b_1)} \cdot \sqrt{\frac{pr}{n}} \sqrt{1 \vee \frac{n}{b_1}} \cdot \sqrt{\eta(\mathcal{S}) \vee \eta(Q_1)} \cdot \|Q_1\|. \quad \left(\text{if } p \geq \frac{\log(n \vee b_1) r}{n} \right)
\end{aligned} \tag{46}$$

The Bernstein inequality now dictates that

$$\begin{aligned}
\|Y_1 - pS_1\| &= \left\| \sum_{i,j} Z_{i,j} \right\| \quad (\text{see (41)}) \\
&\lesssim \alpha \max \left(\log(n \vee b_1) \cdot \beta, \sqrt{\log(n \vee b_1)} \cdot \sigma \right) \quad (\text{see Lemma 5}) \\
&\lesssim \alpha \sqrt{\log(n \vee b_1)} \cdot \sqrt{\frac{rp}{n}} \sqrt{1 \vee \frac{n}{b_1}} \cdot \sqrt{\eta(\mathcal{S}) \vee \eta(Q_1)} \cdot \|Q_1\|, \quad (\text{see (46)})
\end{aligned} \tag{47}$$

except with a probability of at most $e^{-\alpha}$. In particular, suppose that

$$p \gtrsim \alpha^2 \nu(Q_1)^2 \left(1 \vee \frac{n}{b_1} \right) \frac{(\eta(\mathcal{S}) \vee \eta(Q_1)) r \log(n \vee b_1)}{n}, \tag{48}$$

so that (40) holds. Then, by substituting (47) back into (40) and then applying (39), we find that

$$\begin{aligned}
\frac{\|P_{\mathcal{S}^\perp} P_{\hat{\mathcal{S}}_1}\|_F}{\sqrt{r}} &\leq \|P_{\mathcal{S}^\perp} P_{\hat{\mathcal{S}}_1}\| \quad (\text{see (39)}) \\
&\leq \frac{2}{p} \cdot \frac{\|Y_1 - pS_1\|}{\sigma_r(Q_1)} \quad (\text{see (40)}) \\
&\lesssim \alpha \sqrt{\log(n \vee b_1) \cdot \frac{r}{pn} \left(1 \vee \frac{n}{b_1} \right) (\eta(\mathcal{S}) \vee \eta(Q_1)) \cdot \nu(Q_1)} \quad (\text{see (47)}) \\
&=: \delta_1(\nu(Q_1), \eta(Q_1)),
\end{aligned} \tag{49}$$

except with a probability of at most $e^{-\alpha}$, and for fixed Q_1 . In order to remove the conditioning on Q_1 , fix $\nu \geq 1$, $1 \leq \eta_1 \leq \frac{b_1}{r}$, and recall the following inequality for events \mathcal{A} and \mathcal{B} :

$$\begin{aligned} \Pr[\mathcal{A}] &= \Pr[\mathcal{A}|\mathcal{B}] \cdot \Pr[\mathcal{B}] + \Pr[\mathcal{A}|\mathcal{B}^C] \cdot \Pr[\mathcal{B}^C] \\ &\leq \Pr[\mathcal{A}|\mathcal{B}] + \Pr[\mathcal{B}^C]. \end{aligned} \quad (50)$$

Set $\mathcal{Q}_1 = \text{span}(Q_1)$ and let \mathcal{E} be the event where both $\nu(Q_1) \leq \nu$ and $\eta(Q_1) \leq \eta_1$. Thanks to the inequality above, we find that

$$\begin{aligned} &\Pr \left[\left\| \frac{P_{\mathcal{S}^\perp} P_{\widehat{\mathcal{S}}_1}}{\sqrt{r}} \right\|_F \gtrsim \delta_1(\nu, \eta_1) \right] \\ &\leq \Pr \left[\left\| \frac{P_{\mathcal{S}^\perp} P_{\widehat{\mathcal{S}}_1}}{\sqrt{r}} \right\|_F \gtrsim \delta(\nu, \eta_1) \mid \mathcal{E} \right] + \Pr[\mathcal{E}^C] \quad (\text{see (50)}) \\ &\leq e^{-\alpha} + \Pr[\nu(Q_1) > \nu] + \Pr[\eta(Q_1) > \eta_1], \quad (\text{see (49)}) \end{aligned} \quad (51)$$

which completes the proof of Lemma 2.

D Proof of Lemma 3

D.1 Setup and Notation

Let us first simplify our notation for the rest of this section. At the start of the iteration $k \in [2 : K]$, we have available an estimate $\mathcal{S}_o = \widehat{\mathcal{S}}_{k-1}$ of the true subspace \mathcal{S} (with o standing for *old*). In a sense, \mathcal{S}_o is our prior knowledge at this point about \mathcal{S} . To unburden the notation, we write $Q = Q_k$ and so forth, thereby dropping the subscripts. By construction, the columns of the coefficient matrix $Q \in \mathbb{R}^{b \times r}$ are independent copies of the random vector q . Setting $S_Q := S \cdot Q^* \in \mathbb{R}^{n \times b}$, we observe each entry of S_Q independently with a probability of p , collect the measurements in $Y \in \mathbb{R}^{n \times b}$, and let $\Omega \subseteq [1 : n] \times [1 : b]$ denote the corresponding random index set (over which we observe S_Q). We write this as $Y = P_\Omega(S_Q)$, where $P_\Omega(\cdot)$ projects onto the index set Ω .

Given the new measurement block Y , we update our estimate of the subspace \mathcal{S} (from the old $\mathcal{S}_o = \widehat{\mathcal{S}}_{k-1}$ to the new $\mathcal{S}_n = \widehat{\mathcal{S}}_k$) as follows: Calculate the random matrix

$$\mathbb{R}^{n \times b} \ni R = Y + P_{\Omega^C}(O(Y)), \quad (52)$$

where $P_{\Omega^C}(\cdot)$ projects onto the complement of index set Ω , and

$$O(Y) = \left[\cdots S_o \cdot (P_{\omega_j} \cdot S_o)^\dagger \cdot Y[:, j] \cdots \right] \in \mathbb{R}^{n \times b}. \quad (53)$$

As usual, $S_o \in \mathbb{R}^{n \times b}$ is an orthonormal basis for the subspace \mathcal{S}_o . Assume that Q is fixed for now. Let R_r denote a rank- r truncation of R , obtained via SVD. Then, our updated estimate is $\mathcal{S}_n = \text{span}(R_r)$.

To control the (largest) principal angle $\theta_1(\mathcal{S}, \mathcal{S}_n)$, our strategy is to consider R as a perturbed copy of $S_Q = S Q^*$ and, in turn, $\mathcal{S}_n = \text{span}(R_r)$ as a perturbed copy of $\mathcal{S} = \text{span}(S_Q)$. Indeed, an application of Lemma 6 yields that

$$\begin{aligned} \|P_{\mathcal{S}^\perp} P_{\mathcal{S}_n}\|_F &\leq \frac{\|P_{\mathcal{S}^\perp} R\|_F}{\sigma_r(S_Q) - \|R - S_Q\|} \\ &= \frac{\|P_{\mathcal{S}^\perp} R\|_F}{\sigma_r(Q) - \|R - S_Q\|}. \quad (S_Q = S Q^*, \ S^* S = I_r) \end{aligned} \quad (54)$$

We next bound the numerator and denominator in the last line above in separate subsections.

D.2 Bounding the Numerator of (54)

To control the numerator, we begin with some preparation. First, recalling the definition of $O(Y)$ from (53), we observe that

$$\begin{aligned} O(Y) &= O(P_\Omega(S_Q)) \quad (Y = P_\Omega(S_Q)) \\ &= O(P_\Omega(P_{S_o}S_Q)) + O(P_\Omega(P_{S_o^\perp}S_Q)) \quad (\text{linearity}) \\ &= P_{S_o}S_Q + O(P_\Omega(P_{S_o^\perp}S_Q)). \quad (\text{see (53)}) \end{aligned} \quad (55)$$

The above decomposition allows us to rewrite R in (52) as

$$\begin{aligned} R &= Y + P_{\Omega^c}(O(Y)) \quad (\text{see (52)}) \\ &= P_\Omega(S_Q) + P_{\Omega^c}(O(Y)) \quad (Y = P_\Omega(S_Q)) \\ &= P_\Omega(S_Q) + P_{\Omega^c}(P_{S_o}S_Q) + [P_{\Omega^c} \circ O \circ P_\Omega](P_{S_o^\perp}S_Q) \quad (\text{see (55)}, f \circ g(x) := f(g(x))) \\ &= S_Q - P_{\Omega^c}(S_Q) + P_{\Omega^c}(P_{S_o}S_Q) + [P_{\Omega^c} \circ O \circ P_\Omega](P_{S_o^\perp}S_Q) \\ &= S_Q - P_{\Omega^c}(P_{S_o^\perp}S_Q) + [P_{\Omega^c} \circ O \circ P_\Omega](P_{S_o^\perp}S_Q) \\ &= S_Q - P_{\Omega^c}(P_{S_o^\perp}S_Q) + [O \circ P_\Omega](P_{S_o^\perp}S_Q) - [P_\Omega \circ O \circ P_\Omega](P_{S_o^\perp}S_Q), \end{aligned} \quad (56)$$

and, consequently,

$$P_{S^\perp}R = -P_{S^\perp} \cdot P_{\Omega^c}(P_{S_o^\perp}S_Q) + P_{S^\perp} \cdot [O \circ P_\Omega](P_{S_o^\perp}S_Q) - P_{S^\perp} \cdot [P_\Omega \circ O \circ P_\Omega](P_{S_o^\perp}S_Q). \quad (P_{S^\perp}S_Q = P_{S^\perp}SQ^* = 0)$$

In particular, with an application of the triangle inequality, it immediately follows that

$$\|P_{S^\perp}R\|_F \leq \|P_{\Omega^c}(P_{S_o^\perp}S_Q)\|_F + \|P_{S^\perp} \cdot [O \circ P_\Omega](P_{S_o^\perp}S_Q)\|_F + \|[P_\Omega \circ O \circ P_\Omega](P_{S_o^\perp}S_Q)\|_F. \quad (57)$$

We proceed by controlling each norm on the right-hand side above through a series of technical lemmas (proved in Appendices E-G).

Lemma 8. *For $\alpha \geq 1$ and except with a probability of at most $e^{-\alpha}$, it holds that*

$$\|P_{\Omega^c}(P_{S_o^\perp}S_Q)\|_F \lesssim \alpha \max\left(\sqrt{\frac{\eta(\mathcal{Q})r}{b}}, \sqrt{1-p}\right) \|P_{S^\perp}P_{S_o}\|_F \cdot \|Q\|.$$

Lemma 9. *For $\alpha \geq 1$ and except with a probability of at most $b^{1-\alpha}$, it holds that*

$$\|P_{S^\perp} \cdot [O \circ P_\Omega](P_{S_o^\perp}S_Q)\|_F \lesssim \frac{\alpha \log^{\frac{3}{2}}(n \vee b)}{\sqrt{p}} \cdot \|P_{S^\perp}P_{S_o}\|_F \cdot \|P_{S^\perp}P_{S_o}\|_F \cdot \|Q\|,$$

provided that $p \gtrsim \alpha^2 \log^3(n \vee b) \cdot \eta(\mathcal{S}_o)r/n$.

Lemma 10. *For $\alpha \geq 1$ and except with a probability of at most $e^{-\alpha}$, it holds that*

$$\|[P_\Omega \circ O \circ P_\Omega](P_{S_o^\perp}S_Q)\|_F \lesssim \alpha \max\left(\sqrt{\frac{\eta(\mathcal{Q})r}{b}}, \sqrt{1-p}\right) \|P_{S^\perp}P_{S_o}\|_F \cdot \|Q\|.$$

Applying Lemmas 8-10 to (57) yields that

$$\begin{aligned} &\|P_{S^\perp}R\|_F \\ &\leq \|P_{\Omega^c}(P_{S_o^\perp}S_Q)\|_F + \|P_{S^\perp} \cdot [O \circ P_\Omega](P_{S_o^\perp}S_Q)\|_F + \|[P_\Omega \circ O \circ P_\Omega](P_{S_o^\perp}S_Q)\|_F \quad (\text{see (57)}) \\ &\lesssim \alpha \left(\sqrt{\frac{\eta(\mathcal{Q})r}{b}} + \sqrt{1-p} + \frac{\log^{\frac{3}{2}}(n \vee b) \|P_{S^\perp}P_{S_o}\|_F}{\sqrt{p}} \right) \|P_{S^\perp}P_{S_o}\|_F \cdot \|Q\|, \end{aligned} \quad (58)$$

except with a probability of at most $O(e^{-\alpha} + b^{1-\alpha})$, and provided that $p \gtrsim \alpha^2 \log^3(n \vee b) \cdot \eta(\mathcal{S}_o)r/n$.

D.3 Controlling the Denominator of (54)

We now find an upper bound for $\|R - S_Q\|$ in (54). We do so in Appendix H and summarize the outcome below.

Lemma 11. *For $\alpha \geq 1$ and except with a probability of at most $b^{1-\alpha}$, it holds that*

$$\|R - S_Q\| \leq \|R - S_Q\|_F \lesssim \frac{\alpha \log^{\frac{3}{2}}(n \vee b)}{\sqrt{p}} \cdot \|P_{S^\perp} P_{S_o}\|_F \cdot \|Q\|,$$

provided that $p \gtrsim \alpha \log^3(n \vee b) \cdot \eta(S_o)r/n$.

In particular, if we assume that

$$\|P_{S^\perp} P_{S_o}\|_F \lesssim \frac{\sqrt{p}}{\alpha \cdot \nu(Q) \log^{\frac{3}{2}}(n \vee b)}, \quad (59)$$

then Lemma 11 implies that

$$\|R - S_Q\| \leq \frac{\sigma_r(Q)}{2}, \quad (60)$$

with high probability and when p is large enough, as Lemma 11 stipulates.

D.4 Completing the Proof of Lemma 3

For fixed Q , combining (58) and (60) allows us to finally bound the expression in (54) and arrive at

$$\begin{aligned} & \|P_{S^\perp} P_{S_n}\|_F \\ & \leq \frac{\|P_{S^\perp} R\|_F}{\sigma_r(Q) - \|R - S_Q\|} \quad (\text{see (54)}) \\ & \lesssim \alpha \cdot \nu(Q) \left(\sqrt{\frac{\eta(Q)r}{b}} + \sqrt{1-p} + \frac{\log^{\frac{3}{2}}(n \vee b) \|P_{S^\perp} P_{S_o}\|}{\sqrt{p}} \right) \|P_{S^\perp} P_{S_o}\|_F, \end{aligned} \quad (61)$$

except with a probability of at most $O(e^{-\alpha} + b^{1-\alpha})$, under (59), and provided that $p \gtrsim \alpha^2 \log^3(n \vee b) \cdot \eta(S_o)r/n$. If we further assume that

$$\|P_{S^\perp} P_{S_o}\| \leq \frac{\sqrt{p}}{4\alpha \cdot \nu(Q) \log^{\frac{3}{2}}(n \vee b)}, \quad (62)$$

then (61) simplifies to

$$\begin{aligned} \|P_{S^\perp} P_{S_n}\|_F & \lesssim \left(\alpha \cdot \nu(Q) \left(\sqrt{\frac{\eta(Q)r}{b}} + \sqrt{1-p} \right) + \frac{1}{4} \right) \|P_{S^\perp} P_{S_o}\|_F \\ & =: \delta_2(\nu(Q), \eta(Q)). \end{aligned} \quad (63)$$

In summary, conditioned on Q and except with a probability of at most $O(e^{-\alpha} + b^{1-\alpha})$, (63) holds if (59), (62) are met and $p \gtrsim \alpha^2 \log^3(n \vee b) \cdot \eta(S_o)r/n$. In order to remove the conditioning on Q , fix $\nu \geq 1$ and $1 \leq \eta \leq \frac{b}{r}$. Let also \mathcal{E} be the event where both $\nu(Q) \leq \nu$ and $\eta(Q) \leq \eta$. Then, we apply (50) to find that

$$\begin{aligned} & \Pr[\|P_{S^\perp} P_{S_n}\|_F \gtrsim \delta_2(\nu, \eta)] \\ & \leq \Pr[\|P_{S^\perp} P_{S_n}\|_F \gtrsim \delta_2(\nu, \eta) \mid \mathcal{E}] + \Pr[\mathcal{E}^C] \quad (\text{see (50)}) \\ & \leq O(e^{-\alpha} + b^{1-\alpha}) + \Pr[\mathcal{E}^C] \\ & \leq O(e^{-\alpha} + b^{1-\alpha}) + \Pr[\nu(Q) > \nu] + \Pr[\eta(Q) > \eta]. \quad (\text{union bound}) \end{aligned} \quad (64)$$

This completes the proof of Lemma 3.

E Proof of Lemma 8

With Q fixed, note that

$$\begin{aligned}\mathbb{E} [P_{\Omega^C} (P_{\mathcal{S}_o^\perp} S_Q)] &= \mathbb{E} \left[\sum_{i,j} (1 - \epsilon_{i,j}) \cdot (P_{\mathcal{S}_o^\perp} S_Q) [i, j] \cdot E_{i,j} \right] \\ &= (1 - p) \cdot P_{\mathcal{S}_o^\perp} S_Q, \quad (\epsilon_{i,j} \sim \text{Bernoulli}(p))\end{aligned}$$

where $\{\epsilon_{i,j}\}_{i,j} \stackrel{\text{ind.}}{\sim} \text{Bernoulli}(p)$, and $E_{i,j} \in \mathbb{R}^{n \times b}$ is the $[i, j]$ th canonical matrix. We write the deviation from this expectation as

$$\begin{aligned}P_{\Omega^C} (P_{\mathcal{S}_o^\perp} S_Q) - (1 - p) \cdot P_{\mathcal{S}_o^\perp} S_Q &= \sum_{i,j} ((1 - \epsilon_{i,j}) - (1 - p)) \cdot (P_{\mathcal{S}_o^\perp} S_Q) [i, j] \cdot E_{i,j} \\ &=: \sum_{i,j} Z_{i,j},\end{aligned}\tag{65}$$

where $\{Z_{i,j}\} \subset \mathbb{R}^{n \times b}$ are independent and zero-mean random matrices. In order to appeal to the Bernstein inequality, we first compute the β parameter:

$$\begin{aligned}\|Z_{i,j}\|_F &= \|(p - \epsilon_{i,j}) \cdot (P_{\mathcal{S}_o^\perp} S_Q) [i, j] \cdot E_{i,j}\|_F \quad (\text{see (65)}) \\ &= |(p - \epsilon_{i,j}) \cdot (P_{\mathcal{S}_o^\perp} S_Q) [i, j]| \quad (\|E_{i,j}\|_F = 1) \\ &\leq |(P_{\mathcal{S}_o^\perp} S_Q) [i, j]| \quad (\epsilon_{i,j} \in \{0, 1\}) \\ &\leq \|P_{\mathcal{S}_o^\perp} S_Q\|_\infty \\ &\leq \|P_{\mathcal{S}_o^\perp} S\|_{2 \rightarrow \infty} \cdot \|Q\|_{2 \rightarrow \infty} \quad (S_Q = S Q^*, \quad \|AB^*\|_\infty \leq \|A\|_{2 \rightarrow \infty} \cdot \|B\|_{2 \rightarrow \infty}) \\ &\leq \|P_{\mathcal{S}_o^\perp} P_S\|_{2 \rightarrow \infty} \cdot \sqrt{\frac{\eta(Q)r}{b}} \|Q\| \quad (\text{see (7)}) \\ &=: \beta.\end{aligned}\tag{66}$$

Above, $\|A\|_\infty$ and $\|A\|_{2 \rightarrow \infty}$ return the largest entry of A (in magnitude) and the largest ℓ_2 norm of the rows of A , respectively. As for the σ parameter, we write that

$$\begin{aligned}\mathbb{E} \left[\sum_{i,j} \|Z_{i,j}\|_F^2 \right] &= \sum_{i,j} \mathbb{E} \left[(p - \epsilon_{i,j})^2 \cdot |(P_{\mathcal{S}_o^\perp} S_Q) [i, j]|^2 \right] \quad (\text{see (65)}) \\ &= p(1 - p) \cdot \sum_{i,j} |(P_{\mathcal{S}_o^\perp} S_Q) [i, j]|^2 \quad (\epsilon_{i,j} \sim \text{Bernoulli}(p)) \\ &= p(1 - p) \cdot \|P_{\mathcal{S}_o^\perp} S_Q\|_F^2 \\ &\leq (1 - p) \cdot \|P_{\mathcal{S}_o^\perp} S\|_F^2 \cdot \|Q\|^2 \quad (S_Q = S Q^*, \quad \|AB\|_F \leq \|A\|_F \cdot \|B\|) \\ &=: \sigma^2,\end{aligned}\tag{67}$$

and, finally,

$$\max(\beta, \sigma) \lesssim \max \left(\sqrt{\frac{\eta(Q)r}{b}}, \sqrt{1 - p} \right) \cdot \|P_{\mathcal{S}_o^\perp} S\|_F \cdot \|Q\|. \quad (\text{see (66) and (67)})$$

Using the Bernstein inequality and for arbitrary $\alpha \geq 1$, it follows that

$$\begin{aligned}
\|P_{\Omega^C} (P_{\mathcal{S}_o^\perp} S_Q) - (1-p) \cdot P_{\mathcal{S}_o^\perp} S_Q\|_F &= \left\| \sum_{i,j} Z_{i,j} \right\|_F \quad (\text{see (65)}) \\
&\lesssim \alpha \max(\beta, \sigma) \quad (\text{see Lemma 4}) \\
&\lesssim \alpha \max \left(\sqrt{\frac{\eta(\mathcal{Q})r}{b}}, \sqrt{1-p} \right) \cdot \|P_{\mathcal{S}_o^\perp} S\|_F \cdot \|Q\|, \quad (68)
\end{aligned}$$

except with a probability of at most $e^{-\alpha}$. Consequently, with an application of the triangle inequality, we find that

$$\begin{aligned}
\|P_{\Omega^C} (P_{\mathcal{S}_o^\perp} S_Q)\|_F &\lesssim (1-p) \|P_{\mathcal{S}_o^\perp} S_Q\|_F + \alpha \max \left(\sqrt{\frac{\eta(\mathcal{Q})r}{b}}, \sqrt{1-p} \right) \|P_{\mathcal{S}_o^\perp} S\|_F \|Q\| \quad (\text{triangle inequality and (68)}) \\
&\leq (1-p) \|P_{\mathcal{S}_o^\perp} S\|_F \cdot \|Q\| + \alpha \max \left(\sqrt{\frac{\eta(\mathcal{Q})r}{b}}, \sqrt{1-p} \right) \|P_{\mathcal{S}_o^\perp} S\|_F \|Q\| \quad (S_Q = S Q^*) \\
&\lesssim \alpha \max \left(\sqrt{\frac{\eta(\mathcal{Q})r}{b}}, \sqrt{1-p} \right) \|P_{\mathcal{S}_o^\perp} S\|_F \|Q\| \\
&= \alpha \max \left(\sqrt{\frac{\eta(\mathcal{Q})r}{b}}, \sqrt{1-p} \right) \|P_{\mathcal{S}_o^\perp} P_S\|_F \|Q\|, \quad (\text{rotational invariance of Frobenius norm})
\end{aligned}$$

which completes the proof of Lemma 8.

F Proof of Lemma 9

Throughout, we fix the coefficient matrix Q . We begin by noting that

$$\begin{aligned}
\|P_{\mathcal{S}^\perp} \cdot [O \circ P_\Omega] (P_{\mathcal{S}_o^\perp} S_Q)\|_F &= \|P_{\mathcal{S}^\perp} P_{\mathcal{S}_o} \cdot [O \circ P_\Omega] (P_{\mathcal{S}_o^\perp} S_Q)\|_F \quad (\text{see (53)}) \\
&\leq \|P_{\mathcal{S}^\perp} P_{\mathcal{S}_o}\| \cdot \|[O \circ P_\Omega] (P_{\mathcal{S}_o^\perp} S_Q)\|_F \quad (\|AB\|_F \leq \|A\| \cdot \|B\|_F). \quad (69)
\end{aligned}$$

In Appendix I, we bound the random norm in the last line above.

Lemma 12. *For $\alpha \geq 1$ and except with a probability of at most $2b^{1-\alpha}$, it holds that*

$$\|[O \circ P_\Omega] (P_{\mathcal{S}_o^\perp} S_Q)\|_F \lesssim \frac{\alpha \log^{\frac{1}{2}} n \log b}{\sqrt{p}} \cdot \|P_{\mathcal{S}^\perp} P_{\mathcal{S}_o}\|_F \cdot \|Q\|,$$

provided that $p \gtrsim \alpha^2 \log n \log^2 b \cdot \eta(\mathcal{S}_o)r/n$.

We therefore conclude that

$$\begin{aligned}
\|P_{\mathcal{S}^\perp} \cdot [O \circ P_\Omega] (P_{\mathcal{S}_o^\perp} S_Q)\|_F &\leq \|P_{\mathcal{S}^\perp} P_{\mathcal{S}_o}\| \cdot \|[O \circ P_\Omega] (P_{\mathcal{S}_o^\perp} S_Q)\|_F \quad (\text{see (69)}) \\
&\lesssim \frac{\alpha \log^{\frac{1}{2}} n \log b}{\sqrt{p}} \|P_{\mathcal{S}^\perp} P_{\mathcal{S}_o}\| \cdot \|P_{\mathcal{S}^\perp} P_{\mathcal{S}_o}\|_F \cdot \|Q\|, \quad (\text{see Lemma 12}) \quad (70)
\end{aligned}$$

except with a probability of at most $2b^{1-\alpha}$ and provided that $p \gtrsim \alpha^2 \log n \log^2 b \cdot \eta(\mathcal{S}_o)r/n$. This completes the proof of Lemma 9.

G Proof of Lemma 10

Throughout, Q is fixed. Recalling the definition of $O(\cdot)$ from (53), we write that

$$\begin{aligned}
\| [P_\Omega \circ O \circ P_\Omega] (P_{\mathcal{S}_o^\perp} S_Q) \|_F^2 &= \sum_{j=1}^b \left\| (P_{\omega_j} S_o) (P_{\omega_j} S_o)^\dagger \cdot P_{\omega_j} (P_{\mathcal{S}_o^\perp} S \cdot Q[j, :])^* \right\|_2^2 \quad (\text{see (53)}) \\
&= \sum_{j=1}^b \left\| (P_{\omega_j} S_o) (P_{\omega_j} S_o)^\dagger \cdot P_{\mathcal{S}_o^\perp} S \cdot Q[j, :])^* \right\|_2^2 \quad \left((P_{\omega_j} S_o)^\dagger = (S_o^* P_{\omega_j} S_o)^{-1} S_o^* P_{\omega_j} \right) \\
&= \sum_{j=1}^b \| P_{\mathcal{S}_{o,j}} \cdot P_{\mathcal{S}_o^\perp} S \cdot Q[j, :])^* \|_2^2 \quad (\mathcal{S}_{o,j} := \text{span}(P_{\omega_j} S_o)). \tag{71}
\end{aligned}$$

Let also $\mathcal{S}_{o,j}^C := \text{span}(P_{\omega_j^C} S_o)$, and note that $S_o = P_{\omega_j} S_o + P_{\omega_j^C} S_o$ and $(P_{\omega_j} S_o)^* (P_{\omega_j^C} S_o) = 0$. Consequently, $\mathcal{S}_{o,j} \perp \mathcal{S}_{o,j}^C$ and $P_{S_o} = P_{\mathcal{S}_{o,j}} + P_{\mathcal{S}_{o,j}^C}$, from which it follows that

$$\begin{aligned}
\| P_{\mathcal{S}_{o,j}} \cdot P_{\mathcal{S}_o^\perp} S \cdot Q[j, :])^* \| &= \left\| (P_{S_o} - P_{\mathcal{S}_{o,j}^C}) \cdot P_{\mathcal{S}_o^\perp} S \cdot Q[j, :])^* \right\| \\
&= \| P_{\mathcal{S}_{o,j}^C} \cdot P_{\mathcal{S}_o^\perp} S \cdot Q[j, :])^* \|. \tag{72}
\end{aligned}$$

Using the above identity to simplify (71) leads to

$$\begin{aligned}
\| [P_\Omega \circ O \circ P_\Omega] (P_{\mathcal{S}_o^\perp} S_Q) \|_F^2 &= \sum_{j=1}^b \| P_{\mathcal{S}_{o,j}} \cdot P_{\mathcal{S}_o^\perp} S \cdot Q[j, :])^* \|_2^2 \quad (\text{see (71)}) \\
&= \sum_{j=1}^b \| P_{\mathcal{S}_{o,j}^C} \cdot P_{\mathcal{S}_o^\perp} S \cdot Q[j, :])^* \|_2^2 \quad (\text{see (72)}) \\
&= \sum_{j=1}^b \| P_{\mathcal{S}_{o,j}^C} P_{\omega_j^C} \cdot P_{\mathcal{S}_o^\perp} S \cdot Q[j, :])^* \|_2^2 \quad \left(P_{\mathcal{S}_{o,j}^C} = P_{\omega_j^C} S_o (S_o^* P_{\omega_j^C} S_o)^{-1} S_o^* P_{\omega_j^C} \right) \\
&\leq \sum_{j=1}^b \| P_{\omega_j^C} \cdot P_{\mathcal{S}_o^\perp} S \cdot Q[j, :])^* \|_2^2 \\
&= \| P_{\Omega^C} (P_{\mathcal{S}_o^\perp} S Q^*) \|_F^2 \\
&= \| P_{\Omega^C} (P_{\mathcal{S}_o^\perp} S_Q) \|_F^2 \quad (S_Q = S Q^*) \\
&\lesssim \alpha^2 \max \left(\frac{\eta(Q)r}{b}, 1-p \right) \| P_{\mathcal{S}^\perp} P_{S_o} \|_F^2 \cdot \| Q \|^2, \quad (\text{see Lemma 8}) \tag{73}
\end{aligned}$$

except with a probability of at most $e^{-\alpha}$. This completes the proof of Lemma 10.

H Proof of Lemma 11

After recalling (56) and for $\alpha \geq 1$, we observe that

$$\begin{aligned}
& \|R - S_Q\|_F \\
& \leq \left\| -P_{\Omega^c} (P_{\mathcal{S}_o^\perp} S_Q) + [P_{\Omega^c} \circ O \circ P_\Omega] (P_{\mathcal{S}_o^\perp} S_Q) \right\|_F \quad (\text{see (56)}) \\
& \leq \|P_{\mathcal{S}_o^\perp} S_Q\|_F + \|[O \circ P_\Omega] (P_{\mathcal{S}_o^\perp} S_Q)\|_F \quad (\text{triangle inequality, } P_{\Omega^c}(\cdot) \text{ is a non-expansive operator}) \\
& \leq \|P_{\mathcal{S}_o^\perp} S\|_F \cdot \|Q\| + \|[O \circ P_\Omega] (P_{\mathcal{S}_o^\perp} S_Q)\|_F \quad (S_Q = SQ^*, \quad \|AB\|_F \leq \|A\|_F \cdot \|B\|) \\
& \lesssim \|P_{\mathcal{S}_o^\perp} S\|_F \cdot \|Q\| + \frac{\alpha \log^{\frac{1}{2}} n \log b}{\sqrt{p}} \cdot \|P_{\mathcal{S}^\perp} P_{\mathcal{S}_o}\|_F \cdot \|Q\| \quad (\text{see Lemma 12}) \\
& \lesssim \frac{\alpha \log^{\frac{3}{2}}(n \vee b)}{\sqrt{p}} \cdot \|P_{\mathcal{S}^\perp} P_{\mathcal{S}_o}\|_F \cdot \|Q\|,
\end{aligned}$$

except with a probability of at most $b^{1-\alpha}$, and provided that $p \gtrsim \alpha \log^3(n \vee b) \cdot \eta(\mathcal{S}_o) r/n$. This completes the proof of Lemma 11.

I Proof of Lemma 12

Using the definition of $O(\cdot)$ in (53), we write that

$$\begin{aligned}
\|[O \circ P_\Omega] (P_{\mathcal{S}_o^\perp} S_Q)\|_F^2 &= \sum_{j=1}^b \left\| S_o (P_{\omega_j} S_o)^\dagger P_{\omega_j} (P_{\mathcal{S}_o^\perp} S \cdot Q[j, :]^*) \right\|_2^2 \quad ((53) \text{ and } S_Q = SQ^*) \\
&= \sum_{j=1}^b \left\| S_o (P_{\omega_j} S_o)^\dagger (P_{\mathcal{S}_o^\perp} S \cdot Q[j, :]^*) \right\|_2^2 \quad \left((P_{\omega_j} S_o)^\dagger = (S_o^* P_{\omega_j} S_o)^{-1} S_o^* P_{\omega_j} \right) \\
&= \sum_{j=1}^b \left\| (P_{\omega_j} S_o)^\dagger (P_{\mathcal{S}_o^\perp} S \cdot Q[j, :]^*) \right\|_2^2 \quad (S_o^* S_o = I_r)
\end{aligned} \tag{74}$$

For fixed $j \in [1 : b]$, consider the summand in the last line above:

$$\begin{aligned}
\left\| (P_{\omega_j} S_o)^\dagger (P_{\mathcal{S}_o^\perp} S \cdot Q[j, :]^*) \right\|_2 &\leq \left\| (P_{\omega_j} S_o)^\dagger P_{\mathcal{S}_o^\perp} \right\| \cdot \|P_{\mathcal{S}_o^\perp} S \cdot Q[j, :]^*\|_2 \\
&= \left\| (P_{\omega_j} S_o)^\dagger S_o^\perp \right\| \cdot \|P_{\mathcal{S}_o^\perp} S \cdot Q[j, :]^*\|_2 \quad \left(P_{\mathcal{S}_o^\perp} = S_o^\perp (S_o^\perp)^* \right) \\
&=: \|\widehat{Z}_j\| \cdot \|P_{\mathcal{S}_o^\perp} S \cdot Q[j, :]^*\|_2.
\end{aligned} \tag{75}$$

Above, as usual, S_o^\perp is an orthonormal basis for the subspace \mathcal{S}_o^\perp . We can now revisit (74) and write that

$$\begin{aligned}
\|O(P_{\mathcal{S}_o^\perp} S_Q)\|_F^2 &= \sum_{j=1}^b \left\| (P_{\omega_j} S_o)^\dagger (P_{\mathcal{S}_o^\perp} S \cdot Q[j, :]^*) \right\|_2^2 \quad (\text{see (74)}) \\
&\leq \max_j \|\widehat{Z}_j\|^2 \cdot \sum_{j=1}^b \|P_{\mathcal{S}_o^\perp} S \cdot Q[j, :]^*\|_2^2 \quad (\text{see (75)}) \\
&= \max_j \|\widehat{Z}_j\|^2 \cdot \|P_{\mathcal{S}_o^\perp} S Q^*\|_F^2 \\
&\leq \max_j \|\widehat{Z}_j\|^2 \cdot \|P_{\mathcal{S}_o^\perp} S\|_F^2 \|Q\|^2. \quad (\|AB\|_F \leq \|A\|_F \cdot \|B\|)
\end{aligned} \tag{76}$$

It remains to control the maximum in the last line above. First, we focus on controlling $\|\widehat{Z}_j\|$ for fixed $j \in [1 : b]$. Observe that \widehat{Z}_j is a solution of the least-squares problem

$$\widehat{Z}_j := \arg \min_{Z \in \mathbb{R}^{n \times (n-r)}} \|S_o^\perp - (P_{\omega_j} S_o) Z\|_F^2,$$

and, therefore, satisfies the *normal equation*

$$(P_{\omega_j} S_o)^* \left((P_{\omega_j} S_o) \widehat{Z}_j - S_o^\perp \right) = 0,$$

which is itself equivalent to

$$(S_o^* P_{\omega_j} S_o) \widehat{Z}_j = S_o^* P_{\omega_j} S_o^\perp. \quad \left(P_{\omega_j}^2 = P_{\omega_j} \right) \quad (77)$$

In fact, since

$$\begin{aligned} \mathbb{E} [S_o^* P_{\omega_j} S_o^\perp] &= p \cdot S_o^* S_o^\perp = 0, \\ \mathbb{E} [S_o^* P_{\omega_j} S_o] &= p \cdot I_r, \quad (S_o^* S_o = I_r) \end{aligned}$$

we can rewrite (77) as

$$(S_o^* P_{\omega_j} S_o - \mathbb{E} [S_o^* P_{\omega_j} S_o]) \widehat{Z}_j + p \cdot \widehat{Z}_j = S_o^T P_{\omega_j} S_o^\perp - \mathbb{E} [S_o^T P_{\omega_j} S_o^\perp].$$

An application of the triangle inequality immediately implies that

$$p \|\widehat{Z}_j\| \leq \|S_o^* P_{\omega_j} S_o - \mathbb{E} [S_o^* P_{\omega_j} S_o]\| \cdot \|\widehat{Z}_j\| + \|S_o^* P_{\omega_j} S_o^\perp - \mathbb{E} [S_o^* P_{\omega_j} S_o^\perp]\|. \quad (78)$$

To control $\|\widehat{Z}_j\|$, we therefore need to derive large deviation bounds for the two remaining norms on the right-hand side above. For the first spectral norm, we write that

$$\|S_o^* P_{\omega_j} S_o - \mathbb{E} [S_o^* P_{\omega_j} S_o]\| = \left\| \sum_i (\epsilon_i - p) \cdot S_o^* E_{i,i} S_o \right\| =: \left\| \sum_i A_i \right\|, \quad (79)$$

where $\{\epsilon_i\}_i \stackrel{\text{ind.}}{\sim} \text{Bernoulli}(p)$, and $E_{i,i} \in \mathbb{R}^{n \times n}$ is the $[i, i]$ th canonical matrix. Furthermore, $\{A_i\}_i \subset \mathbb{R}^{r \times r}$ are independent and zero-mean random matrices. To apply the Bernstein inequality, we first compute β as

$$\begin{aligned} \|A_i\| &= \|(\epsilon_i - p) \cdot S_o^* E_{i,i} S_o\| \quad (\text{see (79)}) \\ &\leq \|S_o^* E_{i,i} S_o\| \quad (\epsilon_i \in \{0, 1\}) \\ &= \|S_o[i, :]\|_2^2 \\ &\leq \frac{\eta(S_o) r}{n} =: \beta. \quad (\text{see (7)}) \end{aligned} \quad (80)$$

To compute σ , we write that

$$\begin{aligned}
\left\| \mathbb{E} \left[\sum_i A_i^2 \right] \right\| &= \left\| \sum_i \mathbb{E} \left[(\epsilon_i - p)^2 \right] (S_o^* E_{i,i} S_o)^2 \right\| && \text{(see (79))} \\
&= \left\| \sum_i p(1-p) (S_o^* E_{i,i} S_o)^2 \right\| && (\epsilon_i \sim \text{Bernoulli}(p)) \\
&\leq p \left\| \sum_i (S_o^* E_{i,i} S_o)^2 \right\| \\
&= p \left\| \sum_i S_o^* E_{i,i} S_o S_o^* E_{i,i} S_o \right\| \\
&= p \left\| \sum_i \|S_o[i, :]\|_2^2 \cdot S_o^* E_{i,i} S_o \right\| \\
&\leq p \cdot \max_i \|S_o[i, :]\|_2^2 \cdot \left\| \sum_i S_o^* E_{i,i} S_o \right\| \\
&= p \left\| \sum_i S_o^* E_{i,i} S_o \right\| \\
&= p \cdot \frac{\eta(S_o) r}{n} \cdot \left\| \sum_i S_o^* E_{i,i} S_o \right\| && \text{(see (7))} \\
&= p \cdot \frac{\eta(S_o) r}{n} \cdot \left(\sum_i E_{i,i} = I_n, S_o^* S_o = I_r \right) \\
&=: \sigma^2.
\end{aligned} \tag{81}$$

It also follows that

$$\begin{aligned}
\max \left(\log r \cdot \beta, \sqrt{\log r} \cdot \sigma \right) &= \max \left(\frac{\log r \cdot \eta(S_o) r}{n}, \sqrt{\frac{\log r \cdot p \cdot \eta(S_o) r}{n}} \right) && \text{(see (80) and (81))} \\
&\leq \sqrt{\frac{\log r \cdot p \cdot \eta(S_o) r}{n}}. && \left(\text{if } p \geq \frac{\log r \cdot \eta(S_o) r}{n} \right)
\end{aligned} \tag{82}$$

As a result, for $\alpha \geq 1$ and except with a probability of at most $e^{-\alpha}$, it holds that

$$\begin{aligned}
\|S_o^* P_{\omega_j} S_o - \mathbb{E} [S_o^* P_{\omega_j} S_o]\| &\lesssim \alpha \max \left(\log r \cdot \beta, \sqrt{\log r} \cdot \sigma \right) && \text{(see Lemma 5)} \\
&\leq \alpha \sqrt{\frac{\log r \cdot p \cdot \eta(S_o) r}{n}}.
\end{aligned} \tag{83}$$

On the other hand, in order to apply the Bernstein inequality to the second spectral norm in (78), we write that

$$\|S_o^* P_{\omega_j} S_o^\perp - \mathbb{E} [S_o^* P_{\omega_j} S_o^\perp]\| = \left\| \sum_i (\epsilon_i - p) S_o^* E_{i,i} S_o^\perp \right\| =: \left\| \sum_i A_i \right\|, \tag{84}$$

where $\{\epsilon_i\}_i \stackrel{\text{ind.}}{\sim} \text{Bernoulli}(p)$, $E_{i,i} \in \mathbb{R}^{n \times n}$ is the i th canonical matrix, and $\{A_i\}_i \subset \mathbb{R}^{r \times (n-r)}$ are zero-mean and independent random matrices. To compute the β parameter, we write that

$$\begin{aligned}
\|A_i\| &= \|(\epsilon_i - p) S_o^* E_{i,i} S_o^\perp\| \quad (\text{see (84)}) \\
&\leq \|S_o^* E_{i,i} S_o^\perp\| \quad (\epsilon_i \in \{0, 1\}) \\
&\leq \|S_o^* E_{i,i}\| \quad \left((S_o^\perp)^* S_o^\perp = I_{n-r} \right) \\
&= \|S_o[i, :]\|_2 \\
&\leq \sqrt{\frac{\eta(\mathcal{S}_o) r}{n}} =: \beta. \quad (\text{see (7)})
\end{aligned} \tag{85}$$

To compute the σ parameter, we notice that

$$\begin{aligned}
\left\| \mathbb{E} \left[\sum_i A_i A_i^* \right] \right\| &= \left\| \sum_i \mathbb{E} [(\epsilon_i - p)^2] S_o^* E_{i,i} S_o^\perp (S_o^\perp)^* E_{i,i} S_o \right\| \quad (\text{see (84)}) \\
&= \left\| \sum_i p(1-p) \cdot S_o^* E_{i,i} S_o^\perp (S_o^\perp)^* E_{i,i} S_o \right\| \quad (\epsilon_i \sim \text{Bernoulli}(p)) \\
&\leq p \left\| \sum_i S_o^* E_{i,i} S_o^\perp (S_o^\perp)^* E_{i,i} S_o \right\| \\
&\leq p \left\| \sum_i S_o^* E_{i,i} E_{i,i} S_o \right\| \quad \left(S_o^\perp (S_o^\perp)^* \preceq I_n \right) \\
&= p \left\| \sum_i S_o^* E_{i,i} S_o \right\| \\
&= p \cdot \left(\sum_i E_{i,i} = I_n, S_o^* S_o = I_r \right)
\end{aligned} \tag{86}$$

In a similar fashion, we find that

$$\begin{aligned}
\left\| \mathbb{E} \left[\sum_i A_i^* A_i \right] \right\| &\leq p \left\| \sum_i (S_o^\perp)^* E_{i,i} S_o S_o^* E_{i,i} S_o^\perp \right\| \\
&= p \left\| \sum_i \|S_o[i, :]\|_2^2 \cdot (S_o^\perp)^* E_{i,i} S_o^\perp \right\| \\
&\leq p \cdot \max_i \|S_o[i, :]\|_2^2 \cdot \left\| \sum_i (S_o^\perp)^* E_{i,i} S_o^\perp \right\| \\
&= p \cdot \max_i \|S_o[i, :]\|_2^2 \cdot \left(\sum_i E_{i,i} = I_n, (S_o^\perp)^* S_o^\perp = I_{n-r} \right) \\
&= p \cdot \frac{\eta(\mathcal{S}_o) r}{n}, \quad (\text{see (7)})
\end{aligned} \tag{87}$$

and, finally,

$$\begin{aligned}
\sigma &= \max \left(\left\| \mathbb{E} \sum_i A_i A_i^* \right\|, \left\| \mathbb{E} \sum_i A_i^* A_i \right\| \right) \\
&= \max \left(\sqrt{p}, \sqrt{p} \cdot \sqrt{\frac{\eta(\mathcal{S}_o) r}{n}} \right) \quad (\text{see (86) and (87)}) \\
&= \sqrt{p}. \quad \left(\eta(\mathcal{S}_o) \leq \frac{n}{r} \right)
\end{aligned} \tag{88}$$

We now compute

$$\begin{aligned} \max \left(\log n \cdot \beta, \sqrt{\log n} \cdot \sigma \right) &= \max \left(\log n \sqrt{\frac{\eta(\mathcal{S}_o) r}{n}}, \sqrt{\log n \cdot p} \right) \quad (\text{see (85) and (88)}) \\ &= \sqrt{\log n \cdot p}. \quad \left(\text{if } p \geq \frac{\log n \cdot \eta(\mathcal{S}_o) r}{n} \right) \end{aligned} \quad (89)$$

Therefore, for $\alpha \geq 1$ and except with a probability of at most $e^{-\alpha}$, it holds that

$$\begin{aligned} \|S_o^* P_{\omega_j} S_o^\perp - \mathbb{E} [S_o^* P_{\omega_j} S_o^\perp]\| &\lesssim \alpha \max \left(\log n \cdot \beta, \sqrt{\log n} \cdot \sigma \right) \quad (\text{see Lemma 5}) \\ &= \alpha \sqrt{\log n \cdot p}. \end{aligned} \quad (90)$$

Overall, substituting the large deviation bounds (83) and (90) into (78), we find that

$$\begin{aligned} p \|\widehat{Z}_j\| &\leq \|S_o^* P_{\omega_j} S_o - \mathbb{E} [S_o^* P_{\omega_j} S_o]\| \cdot \|\widehat{Z}_j\| + \|S_o^* P_{\omega_j} S_o^\perp - \mathbb{E} [S_o^* P_{\omega_j} S_o^\perp]\| \quad (\text{see (78)}) \\ &\lesssim \alpha \sqrt{\frac{\log r \cdot p \cdot \eta(\mathcal{S}_o) r}{n}} \cdot \|\widehat{Z}_j\| + \alpha \sqrt{\log n \cdot p}, \quad (\text{see (83) and (90)}) \end{aligned}$$

except with a probability of at most $2e^{-\alpha}$ and under (82) and (89). It immediately follows that

$$\begin{aligned} \|\widehat{Z}_j\| &\lesssim \frac{\alpha \sqrt{\frac{\log n}{p}}}{1 - \sqrt{\frac{\alpha^2 \log r \cdot \eta(\mathcal{S}_o) r}{pn}}} \quad (\text{see the next line}) \\ &\lesssim \alpha \sqrt{\frac{\log n}{p}}, \quad \left(\text{if } \frac{\alpha^2 \log r \cdot \eta(\mathcal{S}_o) r}{pn} \lesssim 1 \right) \end{aligned} \quad (91)$$

except with a probability of at most $2e^{-\alpha}$. In light of (82) and (89), we assume that $p \gtrsim \alpha^2 \log n \cdot \eta(\mathcal{S}_o) r/n$. Then, using the union bound and with the choice of $\alpha = \alpha' \log b$, it follows that

$$\max_{j \in [1:b]} \|\widehat{Z}_j\| \lesssim \alpha' \log b \sqrt{\frac{\log n}{p}},$$

provided that $p \gtrsim \alpha'^2 \log^2 b \cdot \log n \cdot \eta(\mathcal{S}_o) r/n$ and except with a probability of at most $2be^{-\alpha' \log b} = 2b^{1-\alpha'}$. Finally, invoking (76), we conclude that

$$\begin{aligned} \|O(P_{\mathcal{S}_o^\perp} S_Q)\|_F &\leq \max_j \|\widehat{Z}_j\| \cdot \|P_{\mathcal{S}_o^\perp} P_S\|_F \|Q\| \quad (\text{see (76)}) \\ &\lesssim \alpha' \log b \sqrt{\frac{\log n}{p}} \cdot \|P_{\mathcal{S}_o^\perp} P_S\|_F \|Q\|, \end{aligned}$$

which completes the proof of Lemma 12.

J Properties of a Standard Random Gaussian Matrix

As a supplement to Remark 4, we show here that a standard random Gaussian matrix $G \in \mathbb{R}^{b \times r}$ is well-conditioned and incoherent, when $b \gtrsim r$. Let $\sigma_{\max}(G) \geq \sigma_{\min}(G)$ denote the largest and smallest singular values of G , respectively. From [29, Corollary 5.35] and for fixed $\alpha \geq 1$, recall that

$$\sqrt{b} - C_3 \alpha \sqrt{r} \leq \sigma_{\min}(G) \leq \sigma_{\max}(G) \leq \sqrt{b} + C_3 \alpha \sqrt{r}, \quad (92)$$

for an absolute constant $C_3 > 0$ and except with a probability of at most $e^{-\alpha^2 r}$. It follows that

$$\nu(G) = \frac{\sigma_{\max}(G)}{\sigma_{\min}(G)} \leq \frac{\sqrt{b} + C_3 \alpha \sqrt{r}}{\sqrt{b} - C_3 \alpha \sqrt{r}}, \quad (93)$$

which, if $\sqrt{b} \geq 2C_3\alpha\sqrt{r}$, reduces to $\nu(G) = O(1)$ with high probability.

For the coherence, note that $G(G^*G)^{-\frac{1}{2}} \in \mathbb{R}^{b \times r}$ is an orthonormal basis for $\text{span}(G)$. Using the definition of coherence, we write that

$$\begin{aligned}
\eta(\text{span}(G)) &= \frac{b}{r} \max_{i \in [1:b]} \left\| G[i, :] (G^*G)^{-\frac{1}{2}} \right\|_2^2 \quad (\text{see (7)}) \\
&\leq \frac{b}{r} \max_i \|G[i, :]\|_2^2 \cdot \left\| (G^*G)^{-\frac{1}{2}} \right\|^2 \\
&= \frac{b}{r} \max_i \|G[i, :]\|_2^2 \cdot (\sigma_{\min}(G))^{-2} \\
&\leq \frac{b}{r} \max_i \|G[i, :]\|_2^2 \cdot \left(\sqrt{b} - C_3\alpha\sqrt{r} \right)^{-2} \quad (\text{see (92)}) \\
&\leq \frac{b}{r} \max_i \|G[i, :]\|_2^2 \cdot \left(\frac{b}{2} - C_3^2\alpha^2r \right)^{-1} \quad \left((a-b)^2 \geq \frac{a^2}{2} - b^2 \right) \\
&\lesssim \frac{b}{r} \max_i \|G[i, :]\|_2^2 \cdot (b - C_4\alpha^2r)^{-1}, \tag{94}
\end{aligned}$$

for a constant $C_4 > 0$ and except with a probability of at most $e^{-\alpha^2r}$. For fixed i , $\|G[i, :]\|_2^2$ is a chi-squared random variable with r degrees of freedom so that

$$\Pr \left[\|G[i, :]\|_2^2 \gtrsim t \cdot r \right] \leq e^{-t}, \tag{95}$$

for any $t \geq 1$. An application of the union bound and the choice of $t = C_5\alpha \log b$ then lead us to

$$\Pr \left[\max_{i \in [1:b]} \|G[i, :]\|_2^2 \gtrsim C_5\alpha \log b \cdot r \right] \leq b^{-(C_5\alpha-1)}. \tag{96}$$

Substituting the bound above back into (94) yields that

$$\begin{aligned}
\eta(\text{span}(G)) &\lesssim \frac{b}{r} \max_i \|G[i, :]\|_2^2 \cdot (b - C_4r)^{-1} \quad (\text{see (94)}) \\
&\lesssim \frac{b}{r} \cdot r\alpha \log b \cdot (b - C_4r)^{-1} \quad (\text{see (96)}) \\
&= \frac{\alpha b \log b}{b - C_4\alpha^2r}, \tag{97}
\end{aligned}$$

except with a probability of at most $e^{-\alpha r} + b^{-(C_5\alpha-1)}$. In particular, when $b \geq 2C_4\alpha^2r$, we find that $\eta(\text{span}(G)) \lesssim \alpha \log b$ with high probability.